AI for Nuclear Physics

FY 2026

February 6, 2026

PI: Robert Edwards

Co-PIs: Steven Goldenberg, Diana McSpadden, Kostas Orginos, Nobuo Sato

Jefferson Lab

U.S. DEPARTMENT of ENERGY

FY25 Milestones completed

Aims	Objective number	Milestone	FY25		/26	
			H1	H2	1	H2
1	LQCD Milestone 1	Collect data for PDF to Ioffe time transformations.				
	LQCD Milestone 2	Design suitable tokenization to support bi-directional PDF to Ioffe time transformer.				
	PDF Milestone 1	Collect all the data available at LHAPDF.				
	PDF Milestone 2	Generate PDF data for all the PDF sets using the LHAPDF interpolation software across a dense grid in x and Q2, along with their corresponding 1-sigma confidence bands.				
	JED Milestone 1	Simulate eP data using Pythia and generate a variety of phase space distributions.				
	JED Milestone 2	Design suitable tokenization scheme that integrates the simulated event-level data and the phase space distributions.				
	PDF/JED/LQCD Milestone 3	Design a suitable tokenization scheme that integrates the info file of the Ioffe time, PDF, and JED datasets.				
	PDF/JED/LQCD Milestone 4	Train and validate the embedding models using the tokenized Ioffe time, PDF, and JED data.				
1	Milestone 1	Explore language foundation models for hadronic physics inverse problems.				
2	Milestone 1	Train the base, open-source LLMs in an unsupervised manner using JLab-specific unlabeled data, specifically, the vector database developed in Aim 1.				
	Milestone 2	Carry out a systematic analysis of the latent space with existing tools.				
3	Milestone 1	For the three use cases, develop the questions for the task-specific fine-tuning.				
	Milestone 2	Tune the LLM and validate the data from Milestone 1.				
	Milestone 3	Evaluate the trained LLM from initial LLM and post-transfer learning. Evaluations will include accuracy, biases, training time, and generalizability.				

Overarching goal:

Project goal - capitalize on AI expertise (LLMs) to maximize utility of JLab's domain data and advance lab mission:

Trained models w NP data are envisioned to provide variety of downstream application including solving inverse problems, classification and regeneration of data

Progress:

Transformed NP data into tokens & assimilated by auto-regressive models preserving fidelity in the inverse transform.

Particle events has been the most challenging and several tokenization schemes have been investigated.

Project course:

Extend & build predictive autoregressive model chains using conditional inputs. Utilize to solve inverse problems for hadronic physics.

Use LQCD methods to improve predictions and provide uncertainty quantification

Structured tokenization for data encoding

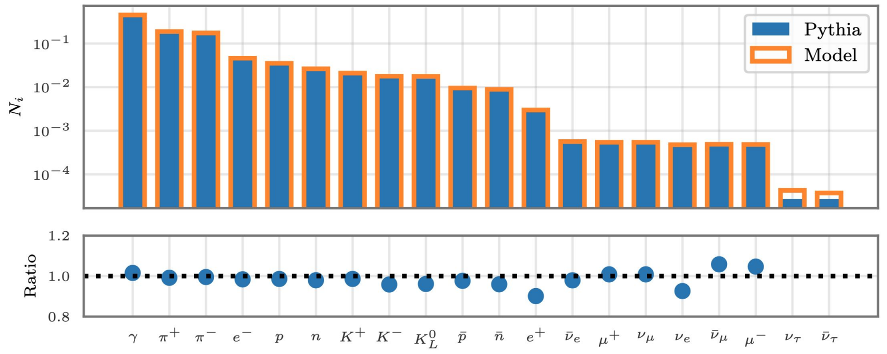
Jefferson Lab

Highlight: Improved representation of e-p data sims at EIC kinematics

- Structuring particle information (pID, transverse momentum ϱ, azimuthal angle η, & pseudorapidity ϕ into multiple tokens):
  - significantly reduced discretization errors compared to current state-of-the-art methods (top)
  - trained auto-regressive transformer-based models to generate events with highly accurate particle fractions (bottom)
- Next step: extend to expt data sets & EIC kinematics

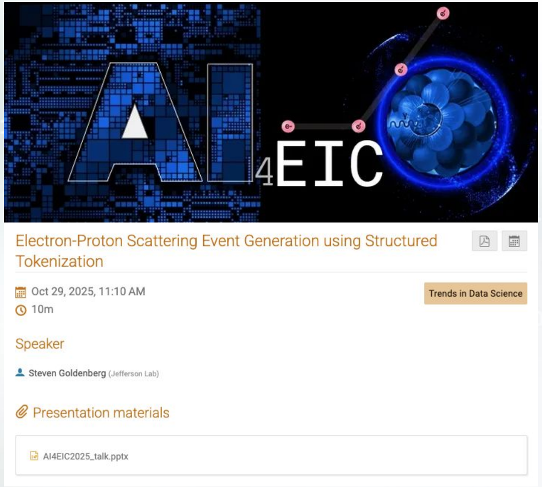
		Codebook		MSE			JS Distance		
	Tokens	Util.	$\|P\|$	p_T	η	ϕ	$\log(p_T)$	η	ϕ
Uniform	20000	23.5%	0.056	2.890	0.336	0.132	0.731	0.728	0.729
	67500	16.0%	0.050	0.907	0.149	0.058	0.690	0.689	0.692
K-means	20000	43.7%	0.173	0.106	0.031	0.030	0.086	0.104	0.145
	67500	31.6%	0.161	0.047	0.013	0.013	0.041	0.058	0.093
VQ-VAE	20000	47.1%	0.164	0.023	0.013	0.016	0.149	0.086	0.142
Structured	1020	**100%**	**0.384**	**1.2e-4**	**3.3e-5**	**1.3e-5**	**0.016**	**0.002**	**0.020**

Table 1: Tokenization metrics on the 8,000 training samples. $\|P\|$ is the normalized perplexity defined by the total codebook perplexity divided by the number of codebook entries. The JS distance is calculated over 100 bins.

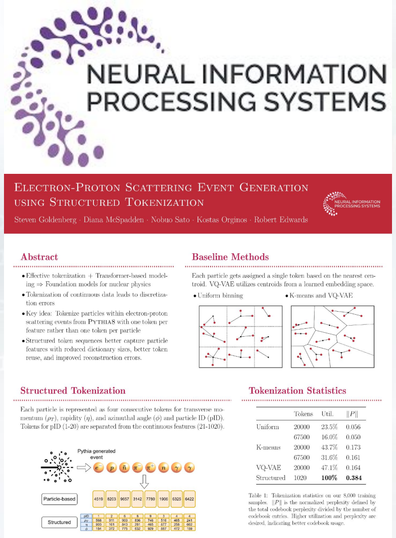


Presentations & publications

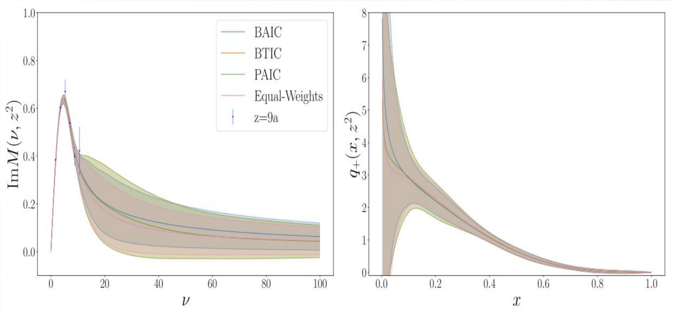
- October: *AI4EIC* - Talk with upcoming submission to *AI4EIC 2025 JINST* Proceedings
- December: *NeurIPS* - Accepted paper and poster presentation for the *Machine Learning and the Physical Sciences* Workshop
- October: *PRD - "Gaussian Process for Inferring Parton Distributions", arXiv:2510.21041*



https://indico.bnl.gov/event/28082/contributions/115796/



https://neurips.cc/virtual/2025/loc/san-diego/122970



https://doi.org/10.48550/arXiv.2510.21041

FY26 plans

Aim 1: World lepton-nucleon(nucleus)

[EIC on a Thumbdrive]

- Improvements and proof-of-concept on other expt data

Aim 2: Event-level inference of hadron structure

- Inverse problems - ensure compatible with DGLAP

Aim 3: Sampling algorithms for lattice field theory

- New scalable/provably correct algorithms for LQCD

FY26 plans change…

Jefferson Lab

Aim 1: World lepton-nucleon(nucleus)

[EIC on a Thumbdrive]

- Improvements and proof-of-concept on other expt data

Aim 2: Event-level inference of hadron structure

- Inverse problems - ensure compatible with DGLAP

Aim 3: Sampling algorithms for lattice field theory

- New scalable/provably correct algorithms for LQCD

U.S. DEPARTMENT of ENERGY | Office of Science

Advanced Scientific Computing Research (ASCR)

THE AMERICAN SCIENCE CLOUD (AmSC)

DOE National Laboratory Program Announcement Number:
LAB 25-3555

Invited Applications: Office of Nuclear Physics (NP) American Science Cloud (AmSC) Data Providers Program (DaPP)

Overview

As a response to Administration and Congressional priorities, working in coordination with the Office of Advanced Scientific Computing Research (ASCR), NP invites labs to submit ambitious pilot applications to advance the NP scientific mission through the development of artificial intelligence (AI) ready datasets and AI models that can take advantage of them. NP is soliciting two-to-three-year proposals for awards up to $5,000,000 for the performance period to provide multiple AI-ready NP datasets and documented expert knowledge that will feed domain-specific AI models. The goal is to use the power of AI to accelerate progress on addressing fundamental questions on the nature of nuclear matter posed in the 2023 Long Range Plan for Nuclear Science: A New Era of Discovery.

- Huge opportunity for JLab
- Large overlaps with lab programs

FY26 Revolution

U.S. DEPARTMENT of ENERGY | Office of Science

Advanced Scientific Computing Research (ASCR)

THE AMERICAN SCIENCE CLOUD (AmSC)

DOE National Laboratory Program Announcement Number:
LAB 25-3555

Invited Applications: Office of Nuclear Physics (NP) American Science Cloud (AmSC) Data Providers Program (DaPP)

Overview

As a response to Administration and Congressional priorities, working in coordination with the Office of Advanced Scientific Computing Research (ASCR), NP invites labs to submit ambitious pilot applications to advance the NP scientific mission through the development of artificial intelligence (AI) ready datasets and AI models that can take advantage of them. NP is soliciting two-to-three-year proposals for awards up to $5,000,000 for the performance period to provide multiple AI-ready NP datasets and documented expert knowledge that will feed domain-specific AI models. The goal is to use the power of AI to accelerate progress on addressing fundamental questions on the nature of nuclear matter posed in the 2023 Long Range Plan for Nuclear Science: A New Era of Discovery.

Huge opportunity - lab awarded multiple projects

AI Ecosystem for NP/HEP [AmSC]

- Build infrastructure to radically reimplement how LQCD and QCD Pheno projects use/scale resources - funded

Model support [ModCon]

- Data entologies (metadata, DOI-s, connectivity) - funded
- Cross-cutting Agentic systems for science - funded

Data prep [DaPP]

- Prepare datasets for consumption by Models
- *Current LDRD fits here - but unfunded under DaPP*

New Directions:

- Integration of theory efforts into AI ecosystem, e.g. chatbots connected to JLab HPC to carryout NP science

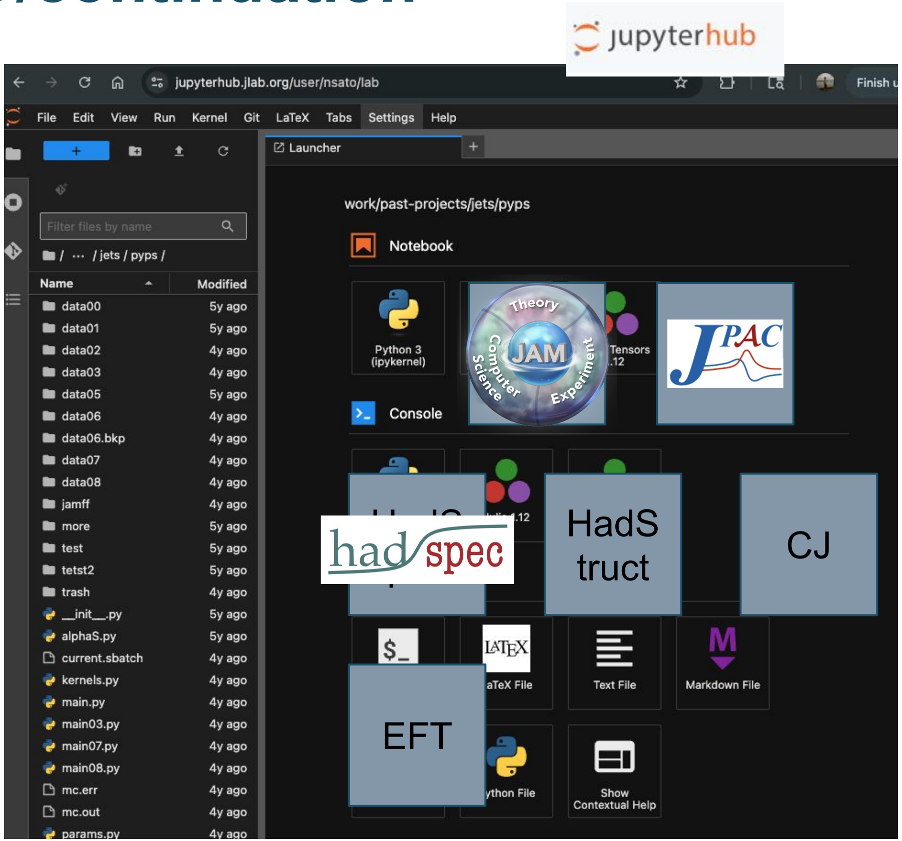
Jefferson Lab

New project direction for LDRD/continuation

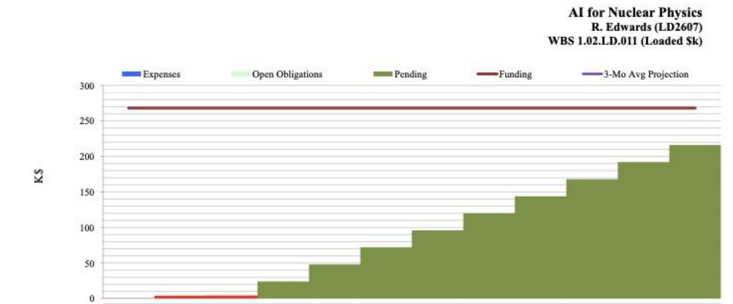
Goal: Consolidate all the theory/numerical tools of the theory division as dedicated AI agentic workflows (curated Chatbots by the theory division)

How:

- Use of Google Co-Scientist for planning the development and testing of the theory agentic workflows
- Use codex(open AI)/vertex (Google) to do the software engineering with dedicated creation of MCP servers monitored by theory staff.
- Deploy the agentic workflows under JLab user accessible space (eg Jupyter hub)



FY26 (& FY27)



	Oct-25	Nov-25	Dec-25	Jan-26	Feb-26	Mar-26	Apr-26	May-26	Jun-26	Jul-26	Aug-26	Sep-26
YTD Spending	0	4	4	24	48	72	96	120	144	168	192	216
Pending	0	0	0	24	48	72	96	120	144	168	192	216
Open Obligations	0	0	0	0	0	0	0	0	0	0	0	0
Expenses	0	0	0	0	0	0	0	0	0	0	0	0
Labor	0	4	4	0	0	0	0	0	0	0	0	0
Funding	268	268	268	268	268	268	268	268	268	268	268	268
3-Mo Avg Projection												

FY26

Redirecting efforts towards harnessing AI ecosystems for Theory applications under AmSC

All projects consistent with the Lab AI strategic direction of ***"femtoscale imaging"***.

Original Budget for FY26

Edwards (Theory): 25% -> 50% AmSC

McSpadden (CST): 30% -> 40% DaPP

Sato (Theory): 5% -> 50% AmSC

Goldenberg (CST): 50% -> 100% (?) ModCon

Karpie (Theory): 25% -> 50% AmSC

Orginos (Theory): 5% -> 0%

Travel: 3K for FY26 - travel used for LDRD

Ask:

Renew project (2nd year) in FY27