Brookhaven National Laboratory

Preparing QCD Data for Foundation Models

September 12, 2025

Lead Applicant/Institution:	Brookhaven National Laboratory
Street Address/City/State/Zip:	Brookhaven Ave, Upton, NY 11973
Postal Address:	P.O. Box 5000, Upton NY 11973
Lead PI:	James C. Dunlop
Partner Laboratories:	Argonne National Laboratory Lawrence Berkeley National Laboratory Oak Ridge National Laboratory Thomas Jefferson National Accelerator Facility
In response to Invitation for Application:	Office of Nuclear Physics (NP) American Science Cloud (AmSC) Data Providers Program (DaPP)

managed by Brookhaven Science Associates
on behalf of the U.S. Department of Energy

U.S. DEPARTMENT of ENERGY

Introduction

Experimental data from RHIC, CEBAF, and heavy-ion collisions at the LHC, together with outputs of Lattice Quantum Chromodynamics (QCD) calculations, form a complex, exabyte-scale resource for scientific discovery. Domain-specific Foundation Models (FMs) have the potential to revolutionize how we access and analyze this vast amount of QCD information by providing general representations of QCD data that can be adapted to multiple downstream applications, such as charged-particle tracking, jet classification, anomaly detection, and surrogate modeling of theoretical predictions. In a recent proof-of-principle study[1,2], a new FM-based tracking algorithm demonstrated scalability and outperformed baseline methods while improving computational efficiency compared to traditional methods.

We propose to extend this concept to the broader landscape of QCD data by curating diverse, multimodal datasets — including experimental measurements from multiple facilities and lattice QCD outputs — into standardized, FAIR[3] and AI-ready, machine-readable formats suitable for FM training. We will work with experimental collaborations and theory groups to establish a unifying data model, develop domain-specific standards for data and metadata, codify expert analysis practices into machine-readable formats, and ensure end-to-end provenance tracking from raw detector signals and lattice configurations through to final physics results. We will also collaboratively establish and implement data governance rules and citability through a persistent identifier architecture. Training and testing datasets will be curated in alignment with FAIR principles and made available to the ASCR AI Consortium and AmSC partners for AI pre-training and fine-tuning, accelerating the development of domain-specific FMs for QCD.

This effort will create a transferable framework for DOE-SC facilities, including the future Electron-Ion Collider, that builds towards the development of AI models that can seamlessly span experimental and theoretical inputs, thus accelerating progress towards the science laid out in the 2023 Long Range Plan for Nuclear Science. We will work in close coordination with aligned proposals in HEP and ASCR to ensure complementarity and impact.

Technical Approach

There are 3 main thrusts to the proposal. The first is to make available data that would allow models such as the FM4NPP proof-of-concept to be extended towards a variety of experiments. We target the computationally intensive processing of experimental data at a nearly raw level, close to the detectors, while recognizing the challenges of interoperability at this level. We intentionally plan to use detectors with different topologies, for instance sPHENIX at RHIC with its many 3-dimensional space points in its Time Projection Chamber (TPC) vs CLAS 12 at CEBAF with its line-like drift chamber response vs purely silicon-based tracking like ATLAS with only a few, high-precision, space points per track. A particularly challenging detector we plan to investigate is the ePIC Barrel Imaging Calorimeter (BIC), which combines full time-based waveforms with precise 3+1+1 dimensional (position + energy + time) images of shower development. We want to scan over a wide range of occupancy, from DIS collisions at CEBAF through to lower multiplicity proton+proton and higher multiplicity Au+Au collisions at RHIC, to the highest multiplicity heavy ion collisions at the LHC in both the Time Projection Chamber in the ALICE experiment and the silicon-based tracking of ATLAS. We want to ensure that

datasets are available over the widest range of conditions encountered in QCD studies to enable the broadest possible investigation of model adaptation to these data.

The second thrust is on theory and simulation, specifically the curation of Lattice QCD (LQCD) data along with accelerated generation of full Monte Carlo events for experimental analysis. We will curate a large collection of gauge configurations obtained using highly improved staggered quarks (HISQ), domain wall fermions (DWF), and clover-improved fermion quarks at zero and non-zero temperature, generated at ANL, BNL, and TJNAF. We will develop workflows to process LQCD data into correlation functions and transform them into the input needed for hadron structure studies at the EIC, along with ontologies for connecting LQCD correlation functions with physics analysis. The overall goal is to provide information in the necessary form to train FMs on these complex datasets.

The third thrust is on the curation of late-stage analysis, to enable the preservation, reproduction, and adjustment of the complex workflows that convert the data into final, publishable physics results. A focus here will be on the engagement with the diverse communities that generate QCD data to agree on common code, data, metadata, and other information needed for FAIR and AI-ready data and workflows. An eventual use-case is to leverage agentic reasoning Foundation Models to provide a distributed search and discover capability such that during the time of the EIC an entry-level graduate student can reproduce an analysis at, say, RHIC with minimal prompts, enabling universality tests such as have been shown to be valuable with reproduction of flow analyses at LEP and HERA as informed by LHC data. More broadly the availability of such preserved workflows will provide a training ground to accelerate autonomous discovery.

Across all three thrusts, and in coordination with the collaborations generating these data, we will focus on provenance collection and data governance throughout the analysis workflow to ensure reproducibility and reuse of the shared data and codes. Datasets provided under this proposal will be fully open for global training thorugh the AmSC infrastructure with appropriate datasheets[4] and model cards[5] and metadata aligned with FAIR principles.

A major role beyond providing the data as algorithmic input is the validation of the algorithmic output. Here the question is the performance of the various models relative to more traditional algorithms for activities such as track reconstruction and the faithfulness of reproduction for lattice QCD data, simulations, and preserved data analyses. As domain experts we plan to provide tools for such validation, potentially even including automation within such tools for additional training.

Project Timeline/Milestones

For the three thrusts, detector-level data (Det), Lattice QCD (LQCD), and Data and Analysis Preservation (DAP), we plan to produce a set of demonstrators on a regular timeframe:

Y1Q1:

- **Det:** Release of simulated 5 million Au+Au sPHENIX TPC data and evaluation chain, enabling FM development for high multiplicity environments. This builds on previous work from the p+p FM dataset publication[6]
- **Det:** Release of similar simulated dataset for Pb+Pb collisions in the ALICE TPC

- **Det:** Release of simulated few million CLAS12 e+P/D data and evaluation chain, enabling FM development for low multiplicity environments relevant for semi-inclusive and exclusive processes
- **DAP:** Engage relevant communities to determine the code, data, metadata, and other information needed for FAIR and AI ready data and workflows

Y1Q2:

- **LQCD:** Curation of large collection (10 PB) of Lattice QCD gauge configurations obtained using clover improved fermion quarks
- **DAP:** Complete data inventory of artifacts for priority datasets and workflows (sPHENIX, STAR, CLAS12, ALICE, Lattice QCD); identify potential AI-ready datasets; and establish prototype catalog infrastructure

Y1Q3:

- **Det:** Release of simulated p+p and Au+Au sPHENIX data and evaluation chain without pileup, along with silicon and calorimeter data. This supports multi-subsystem learning and particle flow jet studies
- **Det:** Release of similar simulated data with ALICE TPC, silicon, and calorimeter
- **Det**: Release of simulated polarized e+p data from CLAS12 and evaluation chain relevant for studying nucleon spin
- **LQCD:** Provide libraries of Lattice QCD gauge configurations at zero and non-zero temperature in international lattice data grid (ILDG) format with additional metadata describing these configurations

Y1Q4:

- **Det:** Release of simulated wave-form-level high dimensional dataset for ePIC Barrel Imaging Calorimeter, including metadata to encode semantic information across its hybrid geometry and an approach for performance validation
- **LQCD:** Provide data on Lattice QCD correlation functions at non-zero temperatures that encode important quantities, such as in-medium quarkonium properties relevant for RHIC and LHC, in AI friendly format using FAIR principles including searchable metadata
- **DAP:** Develop a unified metadata standard for pilot datasets (RHIC + CEBAF). Launch a basic web discovery portal with keyword search and initial provenance capture. Test FAIR curation workflow on selected datasets using AI-assisted metadata enrichment.

Y2Q2:

- **Det:** Release curated real dataset from the Hall C J/psi-007 experiment to demonstrate that our approach can scale from complex multidimensional waveform-level data to a more traditional detector system
- **LQCD:** Provide libraries of raw data files on various two-point and three point-correlation functions needed for hadron structure calculations

for HISQ and DWF formulations in AI friendly format using FAIR principles including searchable metadata

- **DAP:** Implement a federated data location service. Establish a FAIR governance framework with policies, identifiers, and minimal provenance standards. Deploy AI-powered curation assistants to support FAIR compliance for new datasets

Y2Q3:

- **Det:** Release 5 million sPHENIX Au+Au collision events (10 minutes of data), ALICE Pb+Pb collision events, and dataset of e+P/e+D collsion events from CLAS12 for real data applications
- **DAP:** Determine the feasibility of a unified data model across all experiments to enable multi-experiment FMs; and/or harmonizing the outputs for a unified FM trained on reconstructed track-level data

Y2Q4**:**

- **DAP:** Launch an enhanced, federated discovery portal featuring cross-facility search, provenance-aware navigation, and pilot AI-driven recommendations. Add an AI analysis assistant to help physicists update analyses with the latest data, generate code snippets, and document workflows according to FAIR standards
- **LQCD:** Provide workflow and processing tools to obtain correlation functions relevant for specific hadron structure observables of interest at CEBAF and the future EIC

Management Plan and Team Roles

The laboratory teams provide comprehensive and coherent scope across the data sources. BNL as the lead laboratory will coordinate the scope via regular coordination meetings and quarterly milestone reviews. Each institution leads specific deliverables.

ANL: Simulations for ePIC BIC; CEBAF Hall C J/ψ data; Lattice QCD

BNL: RHIC data; Lattice QCD

LBNL: ALICE data

ORNL: Ensuring synergy with ePIC technologies, including streaming

TJNAF: CEBAF data; Lattice QCD

AmSC Integration, Expected Outcomes, and Community Impact

This proposal aims to demonstrate techniques aligned with the AmSC vision, unlocking the exabyte-scale QCD datasets across their full range of modalities for use with modern AI tools. We will work closely with aligned AmSC activities from offices such as ASCR and HEP on standards, to ensure commonality where appropriate and agreed-upon extensions where necessary. The results from this work will provide the foundation for AI-accelerated scientific discovery in support of the 2023 NSAC Long Range plan.

APPENDIX 1: Bibliography

[1] FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics, http://dx.doi.org/10.2139/ssrn.5389206, https://arxiv.org/abs/2508.14087
[2] A Roadmap towards Scaling, Reasoning and Self-evolving Foundation Model for Nuclear and Particle Physics, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5467666
[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[4] Gebru, Timnit et al., Datasheets for datasets, 2021 Association for Computing Machinery, Vol 64 Num 12, DOI: 10.1145/3458723
[5] Mitchell, Margaret et al., Model Cards for Model Reporting, 2019 Association for Computing Machinery DOI: 10.1145/3287560.3287596
[6] TPCpp-10M: Simulated proton-proton collisions in a Time Projection Chamber for AI Foundation Models, https://arxiv.org/abs/2509.05792

APPENDIX 2: Data Management Plan

In this appendix, we describe our plan for complying with the DOE Office of Science requirements to integrate data management planning into the overall research plan for the proposed effort. As required by the Office of Science's Statement on Digital Data Management (http://science.energy.gov/fundingopportunities/digital-data-management/), digital data products generated as part of the proposed research will be preserved for their usability beyond the lifetime of the research activity to enable validation of results and will be distributed openly with a primary focus on sharing with the scientific community to accelerate scientific research. The primary objective of this project is to compile a collection of AI-ready datasets in Nuclear Physics, along with associated models and benchmarks, and make them widely accessible to the AmSC, NP, and scientific AI communities. To achieve this, we will dedicate a significant portion of our resources to acquiring, curating, managing, and distributing high-impact, high-quality datasets that range in size from gigabytes to petabytes. All curated datasets, along with associated models and other software tools, will be made publicly available.

Input Data, Metadata, and Results: If we use any existing data as input, we will clearly document its origin. Input datasets will be cross-referenced or duplicated to improve accessibility, if allowed by the original data source's policy. We will index the datasets we generate so that each release is guaranteed retrievable with a DOI. Any preprints and conference contributions will be posted on arXiv before submission (if permitted by the publisher's policy) and will be updated after publication.

Data Preservation: The long-term preservation of our curated datasets will rely on AmSC infrastructure and policies. During the project's pilot phase, datasets will be stored using the extensive data storage resources available at the participating laboratories. In addition, we will use the publications' archival services to preserve data used to generate charts, figures, or images by submitting it as supplementary information when that mechanism is available.

The supplementary information for our publications will also contain relevant metadata required to replicate the data (e.g., experimental conditions, and computing parameters) that may not be addressed in the original publication. The original publication will provide information regarding how to access the supplementary information.

Confidential and Personal Data: There will be no personal or confidential data used.