Brookhaven National Laboratory

U.S. DEPARTMENT of ENERGY

Treasure: Connections to Nuclear Physics

James Dunlop

April 27, 2026

@BrookhavenLab

NP AmSC Data Providers Program (DaPP)Activity Overview: Preparing QCD Data for Foundation Model

- **Scientific Goals**
  - Curate data from QCD facilities, from detector-level through to final analysis artifacts, along with the output of advanced theoretical calculations, for use by and in developing advanced AI models
- **Significance and Impact**
  - Create transferable framework for SC facilities, including the future Electron Ion Collider, to enable development of AI models that can seamlessly span experimental and theoretical inputs across the pinchpoints of the analysis pipeline
- **Milestones and Figure(s) of Merit**
  - Quarterly releases of AI-ready data in increasing complexity, progressing from simulated to real
  - Progressive development ending in deployment of federated AI discovery portal for analysis artifacts
  - Release of curated AI-ready Lattice QCD data ending with providing workflows for correlation functions
- **Team**
  - BNL – James Dunlop (PI), ANL – Ian Cloet, LBNL – Mateusz Ploskon, ORNL – John Lajoie, TJNAF – Laura Biven

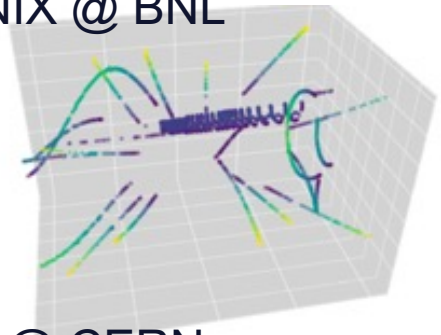
Brookhaven National Laboratory Argonne NATIONAL LABORATORY BERKELEY LAB OAK RIDGE National Laboratory Jefferson Lab

DaPP Overview: Preparing QCD Data for Foundation Models

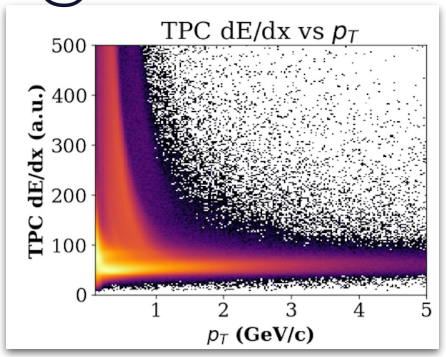
Detector-level Data subproject

- **Scientific Goals**
  - Make available detector simulated and real data at nearly raw level for model training
  - Enable the largest and lowest level NP AI-ready datasets to be used, for example, to train QCD-aware AI foundation models for the most computationally intensive part of the analysis pipeline
- **Significance and Impact**
  - Significant heterogeneity in detector topology and event complexity across experiments complicates model design
  - Release of data over a wide variety of topologies will enable an exploration to what extent models need to be tailored to a given experimental topology vs a more general approach
- **Milestones and Figure(s) of Merit**
  - Y1Q1: Release of simulated TPC-level datasets of Au+Au at sPHENIX and Pb+Pb at ALICE, along with simulated CLAS12 e+P/D data
  - Y1Q3: Release of multi-subsystem (silicon+TPC+calorimeter) simulated data from sPHENIX and ALICE, along with simulated polarized e+p data from CLAS12
  - Y1Q4: Release of simulated wave-form high dimensional data from ePIC Barrel Imaging Calorimeter
  - Y2Q2: Release of curated real dataset from Hall C J/Psi-007 experiment
  - Y2Q3: Release of real collision events from sPHENIX and ALICE, along with CLAS12 e+P/e+D data, for real data application
- **Progress**
  - Data format and data preparation on going

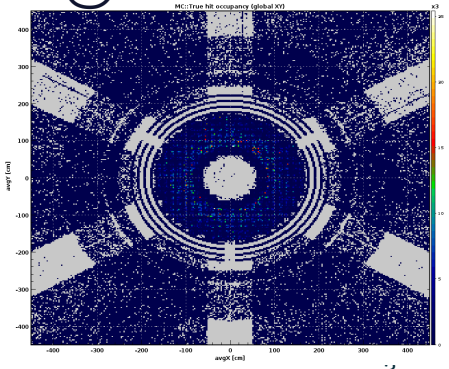
sPHENIX @ BNL

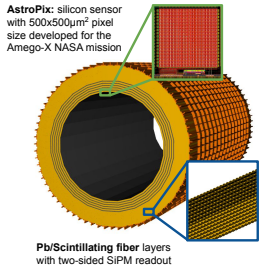


ALICE @ CERN



CLAS12@JLab





ePIC BIC

- FM4NPP, arXiv:2508.14087 **[cs.LG]**, accepted to ICLR26
- Roadmap Whitepaper [ssrn.5467666]
- Open code: https://github.com/FM4NPP
- Open data: [10.5281/zenodo.16970029]

Potential Downstream Model : FM4NPP

A Scaling Foundation Model for Nuclear and Particle Physics

DOE is custodian of exabyte-scale scientific datasets from particle colliders in nuclear and particle physics
Can a self-supervised AI be better at scientific discovery than traditional methods?

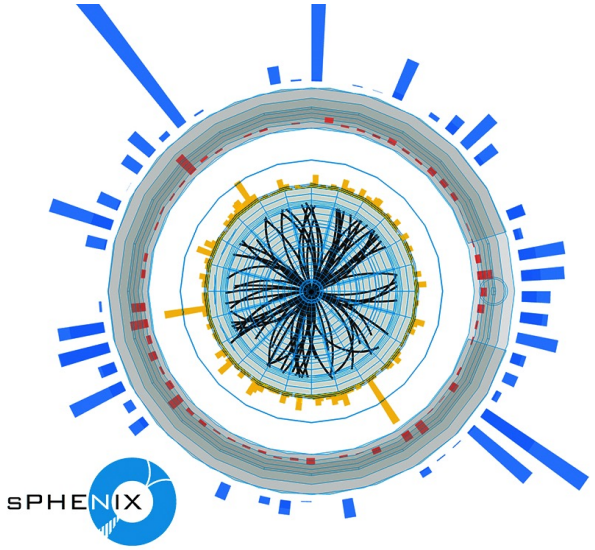
Foundation Model (FM) approach

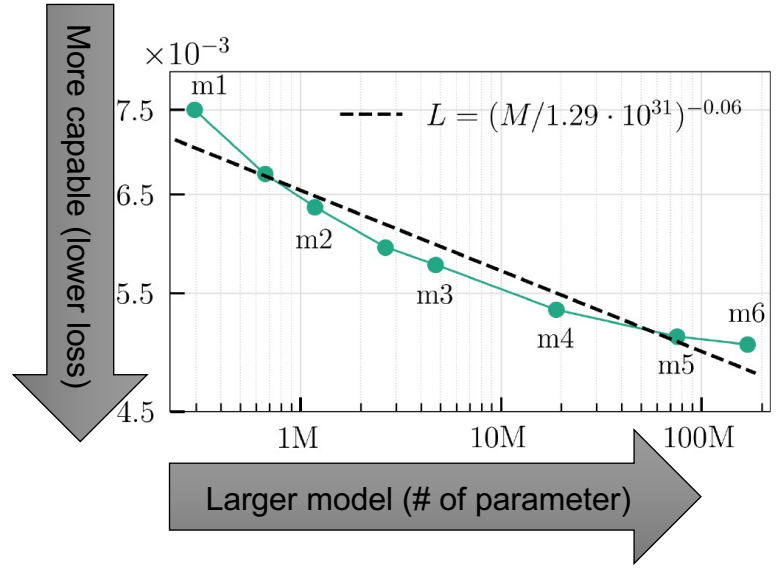
- FM4NPP: a self-supervised FM with unlabeled sparse detector-level data as the input
- Lightweight Adapters connect FM4NPP to multiple downstream tasks → AI Driven Scientific Discovery

Initial investigation established the self-supervised loss scaling behavior of FM4NPP → first in the field
Demonstrated scaling behavior and superior performance for downstream tasks

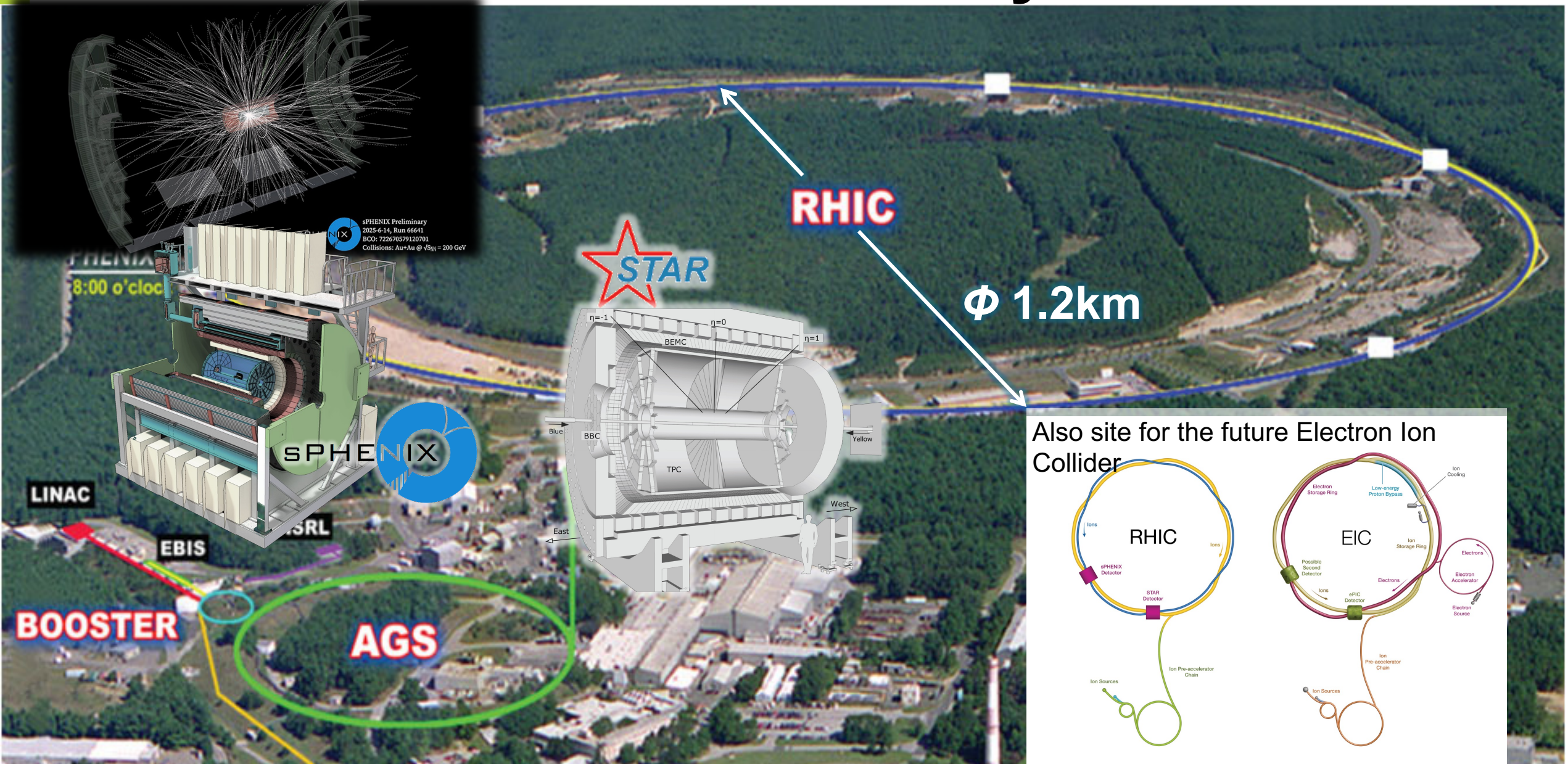
Next phase: scaling to larger and more capable model under Genesis Mission

Large scientific facility (RHIC) in NY recreate particle soup existed in early universe, captured by the three-story high sPHENIX experiment. And its data used to drive this AI research



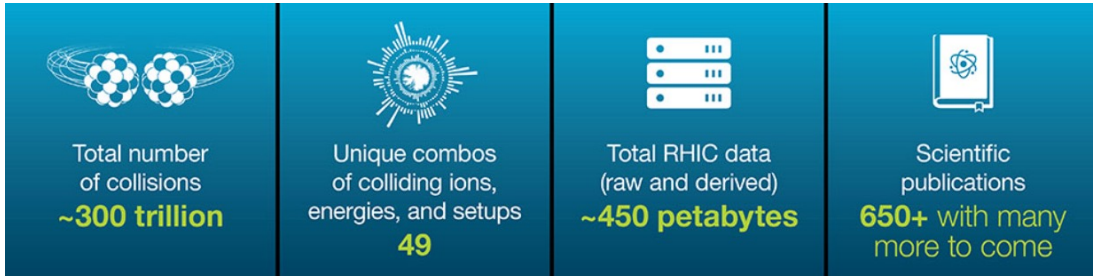


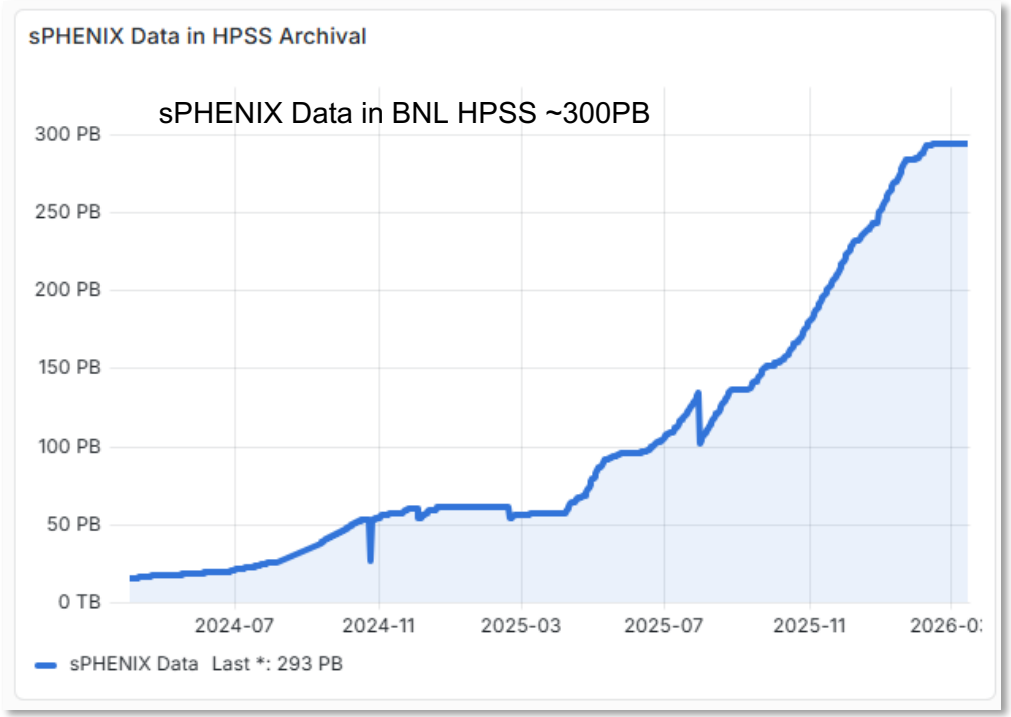
BNL DOE NP Collider Facility and Dataset



BNL DOE NP collider facility and dataset

- BNL hosts the dataset for RHIC and future EIC
  - US largest HPSS storage system at SCDF
- 450 PB data, within which 300PB are from sPHENIX taken in past 3 years [BNL news]
  - An optimal dataset for AI exploration without traditional trigger or filtering bias for QCD emergent phenomenon research
  - **Testing ground for what we will see from the EIC**



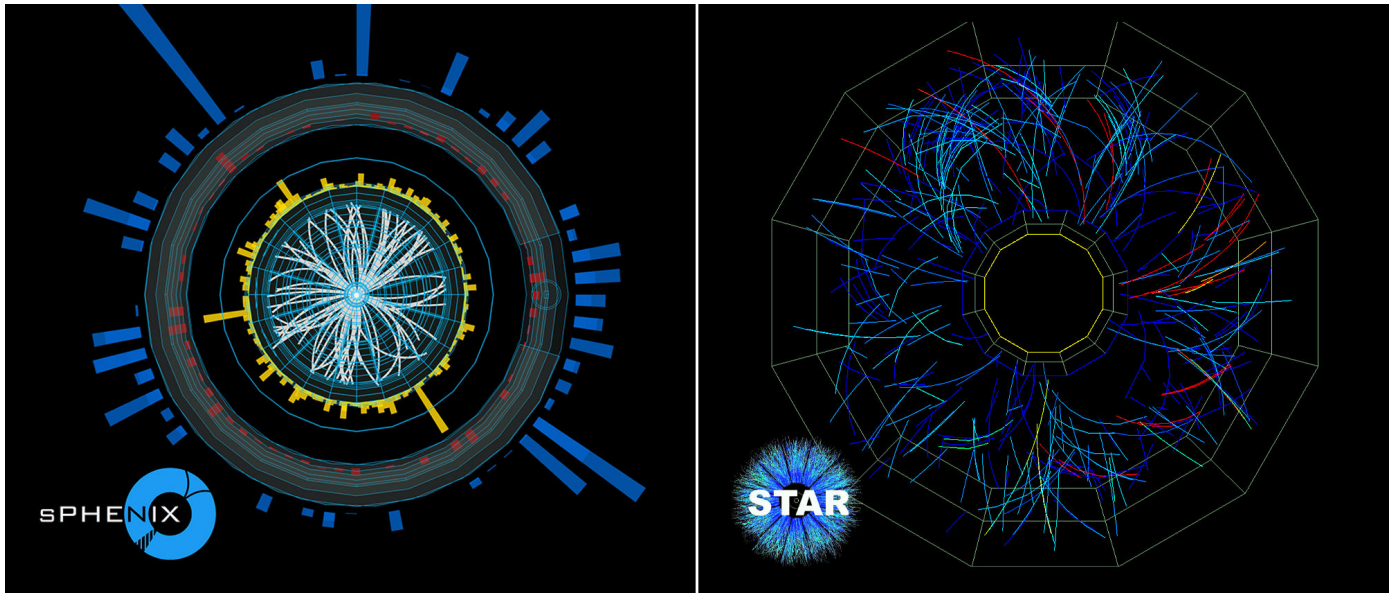


DaPP Overview: Preparing QCD Data for Foundation Models

Data and Analysis Preservation subproject

- **Scientific Goals**
  - Develop a unified, flexible metadata and knowledge framework for QCD experimental data.
  - Capture essential expert knowledge and documentation from legacy and current NP experiments.
  - Enable high-quality NP AI-ready datasets to be used, for example, to train QCD-aware AI foundation models.
- **Significance and Impact**
  - Significant heterogeneity in workflows, data formats, analysis methods, and documentation practices across experiments poses major challenges to create AI-ready datasets.
  - Without coordinated metadata collection, tacit expert knowledge risks being lost as collaborations wind down, reducing long-term scientific value.
  - A standardized, well-structured dataset ecosystem enables reliable QCD-aware AI models, supports cross-experiment analysis, and provides a reusable template for future DOE-SC data initiatives.
- **Milestones and Figure(s) of Merit**
  - Y1Q1: Engage communities to identify code, data, and metadata needs for FAIR and AI-ready workflows.
  - Y1Q2: Complete data inventory of artifacts for priority datasets (sPHENIX, STAR, CLAS12, ALICE, LQCD)
  - Y1Q4: Develop unified metadata standard; launch web discovery portal with AI-assisted curation
  - Y2Q2: Implement federated data location service and FAIR governance framework
  - Y2Q3: Assess feasibility of a unified data model across experiments for multi-experiment FMs
  - Y2Q4: Launch enhanced discovery portal featuring an AI analysis assistant and provenance tracking.

Curation of experimental legacy



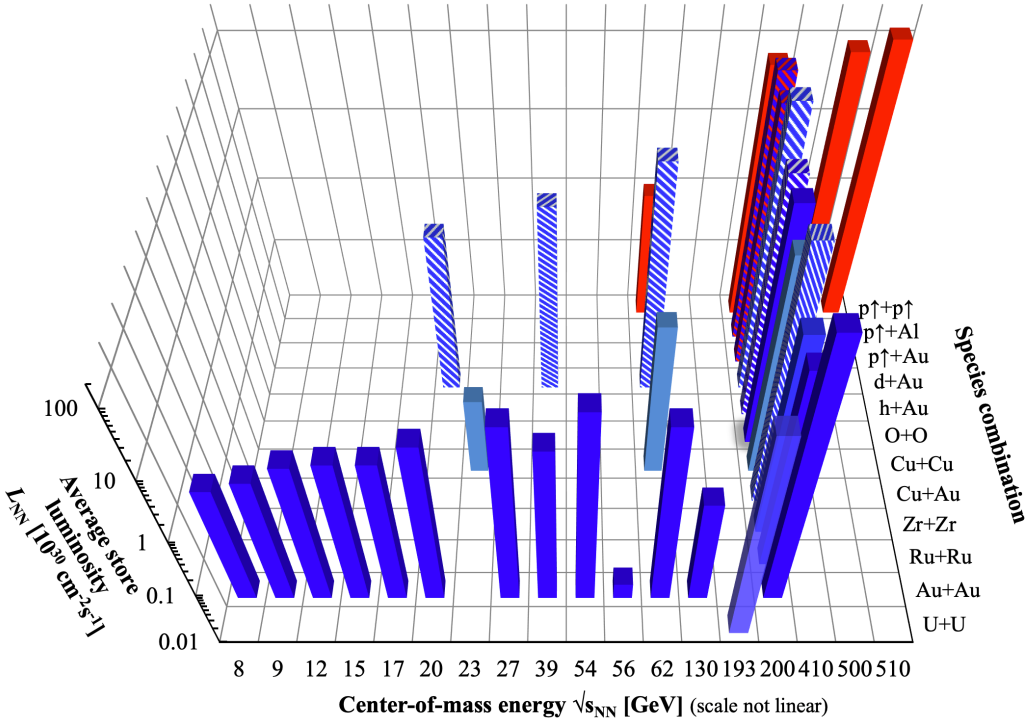
Last collisions at RHIC 9 am February 6, 2026

49 unique combinations, never to be run again, and 25 years of successive upgrades

Curation of unique datasets critical

Fixed-target program at CEBAF also heterogeneous: generic feature of NP datasets

RHIC energies, species combinations and luminosities (Run-1 to 24)



DaPP Activity: Preparing QCD Data for Foundation Models

Lattice QCD subproject

- **Scientific Goals**
  - Curation of Lattice QCD data, with a large collection of gauge configurations using highly improved staggered quarks (HISQ), domain wall fermions (DWF), and clover-improved fermion quarks at zero and non-zero temperature
  - Develop workflows to process LQCD data into correlation functions and transform them into the input needed for hadron structures at the EIC, along with ontologies for connecting LQCD correlation functions with physics analysis
- **Significance and Impact**
  - Enable training of Foundation Models on complex lattice data
- **Milestones and Figure(s) of Merit**
  - Y1Q2: Curation of large collection (10 PB) of Lattice QCD gauge configurations using clover improved fermion quarks
  - Y1Q3: Provide libraries of Lattice QCD gauge configurations at zero and non-zero temperature in ILDG format with additional metadata describing configurations
  - Y1Q4: Provide data on Lattice QCD correlation functions at non-zero temperature, such as in-medium quarkonium properties relevant for RHIC and LHC, in AI friendly format using FAIR principles
  - Y2Q2: Provide libraries of raw data files on various 2-point and 3-point correlation functions needed for hadron structure calculations
  - Y2Q4: Provide workflow and processing tools to obtain correlation functions relevant for specific hadron structure observables of interest at CEBAF and the EIC

Summary

Two-year project QCD Data for Foundation Models to prepare data from multiple facilities and at both ends of the analysis pipeline: raw and final analysis and theory

As broad as possible experimental topologies and analysis techniques, to enable exploration of where common solutions do and don’t work

Multiple points to work together