GENESIS Phase I Proposal

Extending FM4NPP Across Silicon Trackers and Lepton Collider Data

MIT-led coordinated proposal within the broader FM4NPP GENESIS portfolio

Participating institutions: MIT, Iowa State University, and Brookhaven National Laboratory

Focus Area 14A: Foundation Models of Particle Interactions and Cosmic Physics

Collaborators and Coordination

MIT (lead institution)

- Gunther Roland (MIT lead), GM Innocenti, Christof Roland, Hao Ren Jheng, Alex Patton

- **Iowa State University**
  - Marzia Rosati; Cameron Dean (moving from MIT to Iowa State as Assistant Professor)
- **Brookhaven National Laboratory and FM4NPP collaborators**
  - Yihui Ren, Joseph Osborn, Elizabeth Brost, Haiwang Yu, Syed Haider Abidi, Yeonju Go, Peter Boyle, Jin Huang, Alexei Klimentov, Michael Begel, Xin Qian, Torre Wenaus, Dmitri Denisov, Abhay Deshpande, Meifeng Lin, Shinjae Yoo, Viviana Cavaliere, James C. Dunlop, Hong Ma
- **Additional collaborators highlighted in the FM4NPP phase plan**
  - Ron Soltz (LLNL), William Zajc (Columbia), Yeonju Go / Stony Brook collaborators, with broader Phase II scaling partnerships discussed across LLNL, ORNL, LBNL, JLab, ANL, Google, and NVIDIA.
- **Context: one of three coordinated FM4NPP-based GENESIS Phase I efforts intended to merge into a stronger shared GENESIS Phase II proposal.**

FM4NPP Goals and Current Status

FM4NPP is already an active effort: model, datasets, benchmarks, and scaling roadmap are in place; shovel-ready for Genesis

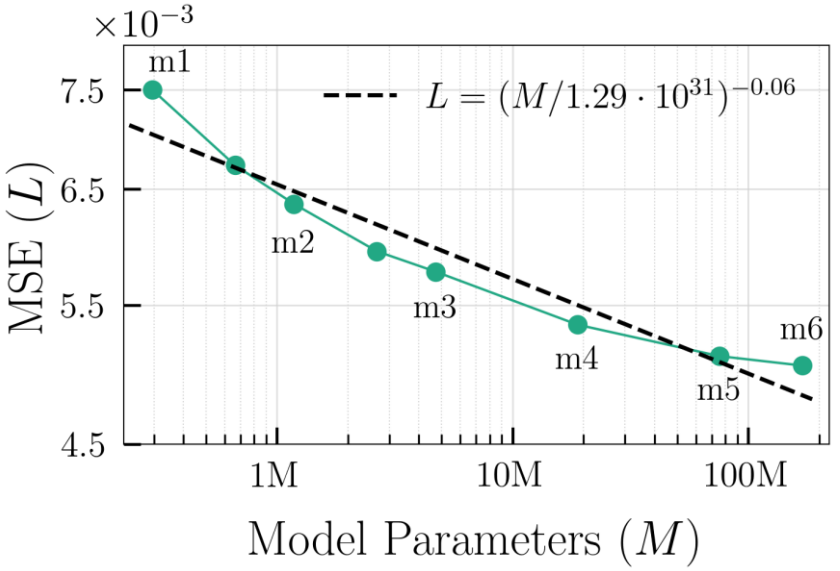
FM4NPP shows clear self-supervised scaling behavior with model size, token count, and compute O(10^7) sPHENIX proton-proton collision events

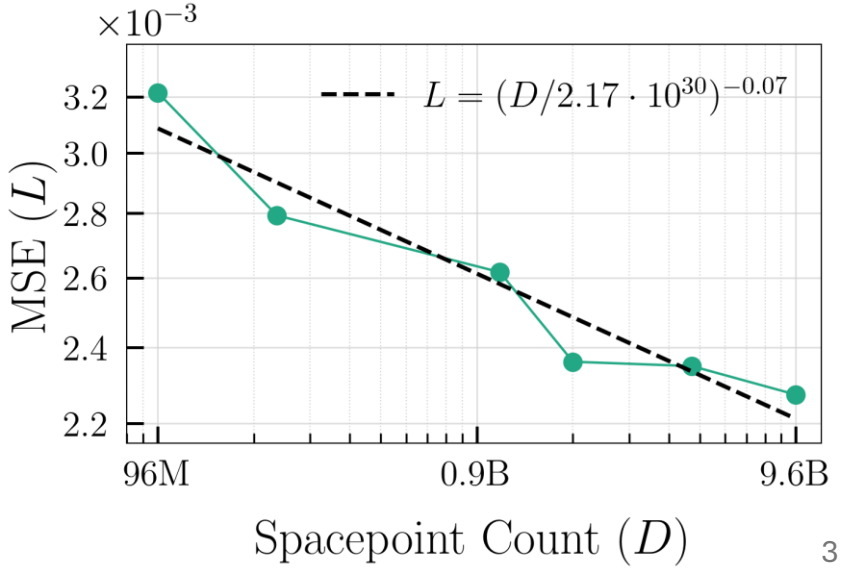
Frozen FM backbones plus lightweight adapters outperform baseline models on downstream tasks, demonstrating transferable representations rather than task-specific memorization

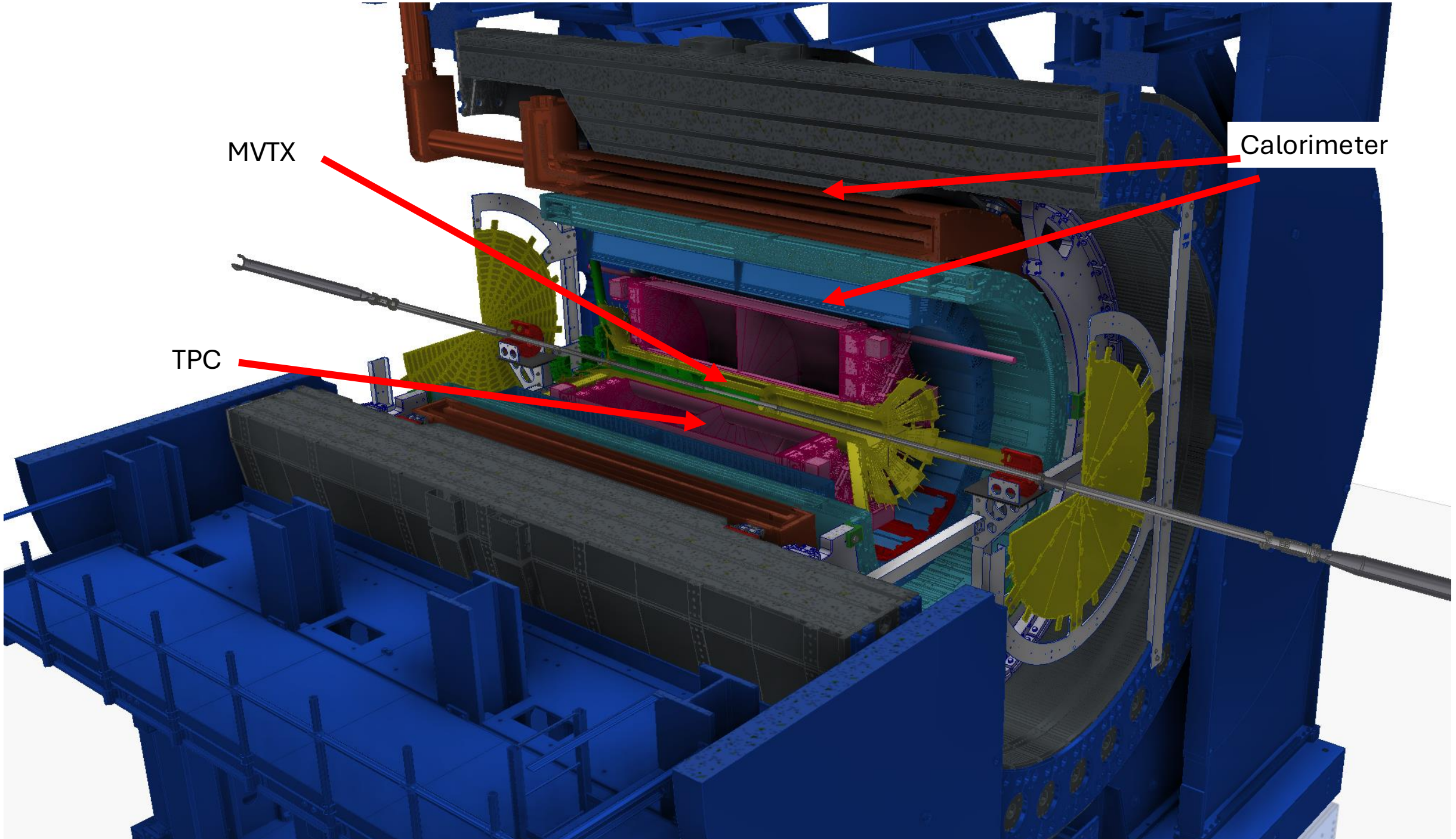
Learned representations are broadly task-agnostic and data-efficient, which is exactly what makes multi-detector and multi-facility expansion plausible

Roadmap exists for real-data integration, multi-detector coverage, and multi-facility transfer. This proposal is the ready-now GENESIS step that adds MVTX plus LEP/ZEUS data to that trajectory

arXiv 2508.14087: https://arxiv.org/pdf/2508.14087 ; SSRN 5467666: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5467666






MVTX
Calorimeter
TPC

Project Overview

Vision: extend FM4NPP beyond its current sPHENIX TPC foundation to learn unified detector representations across silicon tracker data and qualitatively different collision environments.

- Direction 1: incorporate high-granularity silicon-tracker information from detectors such as the sPHENIX MVTX so the model can learn local hit structure, short track segments, and subsystem-complementary tracking information
  - MIT team has leading role in MVTX reconstruction, readout and sPHENIX track seeding, reconstruction. Prior DOE funded effort developing demonstrator for Fast-ML MVTX track reconstruction
- Direction 2: incorporate LEP e+e- detector data from ALEPH, DELPHI, and OPAL, together with ZEUS ep data from HERA, to test cross-facility and cross-interaction transfer in a controlled legacy-data setting
  - Close connection to YJ Lee's electron-positron "recycling frontier" leading effort; synergies with Lee's 14.C agentic AI proposal
- Why AI is essential: the relevant data are heterogeneous in geometry, occupancy, modality, and physics content. A self-supervised foundation model is a natural way to capture shared latent structure across these domains
- Strategic role: ZEUS provides a particularly valuable bridge toward future Electron-Ion Collider analyses, where ep-like detector and event structures can seed the next stage of FM4NPP development

14.A Alignment

Develop and curate the **essential data of nuclear and particle experimental** efforts, critical to **train foundation models** of particle interactions and cosmology to accelerate new breakthroughs in our understanding of the universe. **Data and models may include the future Electron-Ion Collider**, cosmic observations, underground and accelerator-based experiments as well as synthesizing **different modalities of data** from across multiple large-scale sky surveys to understand nuclear astrophysics, dark energy, dark matter, and the physics of the early Universe. Successful scope will seamlessly span experimental and theoretical inputs **across the pinch points of analysis pipelines** from detector-level through to final scientific artifacts, along with the output of advanced theoretical calculations. **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.**

14.A Alignment and Data Pipeline

This proposal directly aligns with Genesis Focus Area 14.A by curating and learning from essential nuclear and particle experimental data to train foundation models of particle interactions.

- It addresses the RFA emphasis on spanning the pinch points of analysis pipelines from detector-level inputs through derived scientific artifacts, while connecting experimental signals to theoretical knowledge.
- It also targets core 14.A technical challenges: sparse-data domains, heterogeneous detector modalities, high-dimensional and petabyte-scale workflows, and scalability of both training and downstream adaptation.
- Phase I data pipeline: curate TPC + MVTX data and harmonize LEP/ZEUS legacy formats -> build common event representations and metadata -> pretrain on mixed-facility corpora -> evaluate adapter-based downstream tasks and transfer -> prepare coordinated Phase II roadmap.
- This scope is especially well matched to 14.A because it synthesizes different detector modalities and collision types while laying groundwork for future EIC-relevant foundation models.

9-months work plan

Months 1-3: Data curation and baselines

Assemble a reproducible corpus spanning sPHENIX TPC + MVTX data and legacy LEP/ZEUS event formats; define a common representation strategy, metadata schema, and benchmark splits.
Establish baseline tracking or event-understanding tasks for each data family to enable comparison against pretrained FM adapters.

Months 4-6: AI-advantage demonstration

Pretrain an updated FM4NPP model on the mixed corpus and adapt it with lightweight heads to silicon-enhanced tracking and cross-facility representation tasks.
Quantify detector transferability, data efficiency, and gains from mixed-domain pretraining at the 6-month go/no-go review.

Months 7-9: Cross-detector transfer and Phase II integration

Run focused transfer studies across TPC-to-MVTX, hadronic-to-e+e-, and hadronic-to-ep settings, using ZEUS as the explicit bridge toward future EIC-style workflows.
Package datasets, checkpoints, and benchmark scripts for coordinated release and merge this effort with the other two FM4NPP GENESIS strands into a unified Phase II proposal.

9-months deliverables & Metrics

- **Deliverables**
  - D1: Curated multimodal Phase I corpus covering sPHENIX TPC + MVTX and legacy LEP/ZEUS detector data, with shared preprocessing and benchmark definitions.
  - D2: Updated FM4NPP checkpoint(s) and adapter recipes for silicon-enhanced tracking and cross-facility event representations.
  - D3: Transfer-study report and a coordinated Phase II roadmap integrated with the other FM4NPP GENESIS proposals.
- **Go/No-Go metrics for AI advantage**
  - M1: Mixed-domain pretraining improves at least one silicon-tracker and one cross-facility benchmark relative to single-domain training.
  - M2: Pretraining reduces labeled-data requirements by at least 4x for selected downstream tasks relative to training task models from scratch.
  - M3: Adding MVTX and LEP/ZEUS data yields measurable transfer gains over the current TPC-only FM4NPP baseline.
  - M4: ZEUS-pretrained representations show promising transfer to EIC-motivated proxy tasks, justifying the joint Phase II expansion.

Workforce Development Plan

- **Graduate students, postdocs, and research staff**
  - At MIT, Phase I will support partial effort for 2 graduate students, 1 postdoc, and 1 research scientist, embedded in a broader cross-institution FM4NPP training environment.
  - Trainees across MIT, Iowa State, and BNL will work at the interface of detector physics, scientific machine learning, and AI-ready data curation for nuclear and particle physics. This will be done within the broader FM4NPP eco system.
- **Concrete work-package ownership**
  - Team members will own specific packages such as MVTX/TPC harmonization, LEP and ZEUS data ingestion, benchmark design, adapter evaluation, and reproducible software release.
- **Undergraduate on-ramp and mentoring**
  - Short focused projects in data conversion, quality checks, stress-testing, and validation will create an accessible undergraduate entry point, supported by joint mentoring, regular technical check-ins, and code review.
- **Broader FM4NPP and agentic AI connection**
  - This workforce plan connects naturally to Yen-Jie Lee's GENESIS agentic AI effort, especially around e+e- archival data, where curated legacy datasets, benchmark tasks, and reusable tools can be shared across the coordinated FM4NPP program.
- **Open science and Phase II readiness**
  - The project will produce reusable schemas, software, checkpoints, and onboarding documentation that lower the barrier for future FM4NPP efforts and prepare trainees for the merged GENESIS Phase II program.