DOE Genesis Mission · Phase I ● Focus area 14c · Expedited Discovery from Petabyte-Scale Datasets

Expedited Discovery by Leveraging Agentic Workflows for the Mu2e Charged Lepton Flavor Violation Search

Yuri Oksuzian for the Mu2e Genesis team

Lead PI D. J. Tedeschi (University of South Carolina) · Argonne · Caltech · Fermilab · University of Virginia · Yale University

Faster optimization
weeks → days
Run-1 re-optimization turnaround

Faster diagnosis
hours → min
Anomaly diagnosis during the Cosmic Ray Run

Better sensitivity
10× larger search
More configurations evaluated; better $R_{\mu e}$ sensitivity than the hand-tuned baseline

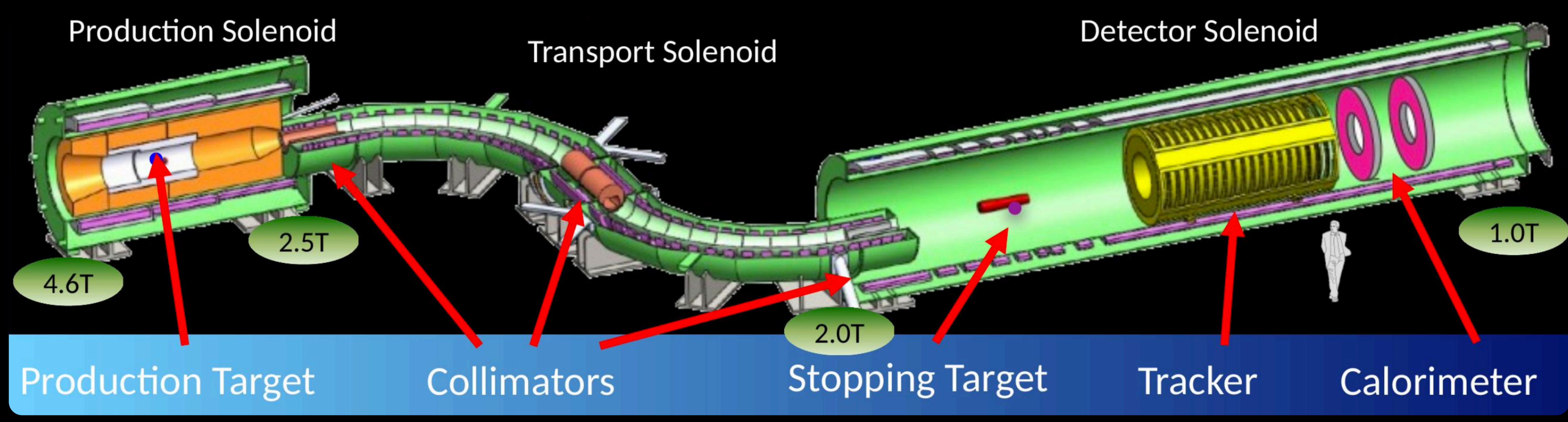
Better data quality
every run checked
Anomalies that shifters would miss are caught in every dataset

Physics case

Mu2e: a 10,000× more sensitive search for $\mu^- N \to e^- N$

Neutrinoless muon-to-electron conversion in the field of an aluminum nucleus violates charged lepton flavor conservation. The $\mu \to e$ conversion rate within the Standard Model is unmeasurably small, so an observation of the 105 MeV conversion electron would be a clear sign of New Physics, sensitive to mass scales up to 10^4 TeV.

Sensitivity depends on experiment design, background control and data quality. This proposal addresses all three.



Approach

Mu2e at Fermilab is in detector commissioning and transitions to a **Cosmic Ray Run (FY27)**, followed by first physics data-taking in **Run-1 (Q1 FY28)**. Both phases give us real data and real operations on which to develop AI-assisted design and monitoring.

We build on two existing DOE HEP AI initiatives and deploy them on an operating Intensity Frontier experiment:

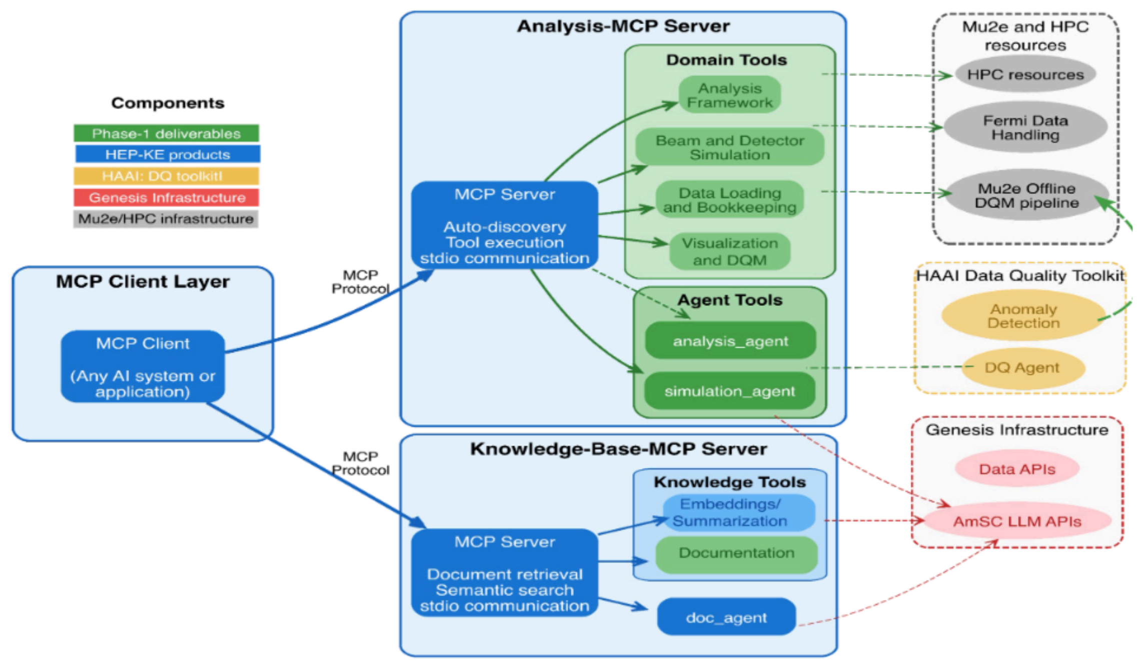
AmSC Knowledge Extraction (KE) provides the orchestration framework and the connection to the American Science Cloud effort. Mu2e software is exposed through Model Context Protocol (MCP) servers.

Hardware-Aware AI (HAAI) provides the data-quality toolkit: transformer-based anomaly detection (TranAD, DeepHydra) and an agent that correlates detector data with operational logs.

Architecture: four MCP interfaces wrap existing physics software

Existing code is not replaced; it is wrapped in an interface layer grounded in experiment documentation. An agent receives a task in natural language, breaks it into tool calls, evaluates the results and adapts its plan, iterating, branching or escalating to a human. All tool calls and reasoning traces are logged so every result can be reproduced.

1. **Simulation.** Geant4 beam transport and detector response; parameter configuration and batch submission on demand.
2. **Analysis.** Physics sensitivity extraction and DQM histogram generation on FermiGrid, as callable tools. Figures of merit are multi-objective, e.g. sensitivity vs. detector radiation damage.
3. **DQM.** Access to the Offline DQM pipeline, HAAI anomaly scores, and LLM-vision interpretation of diagnostic plots.
4. **Documentation.** Semantic search over the Mu2e doc database, wikis, electronic logbooks, run-log and configuration databases. A proof of concept is already running.



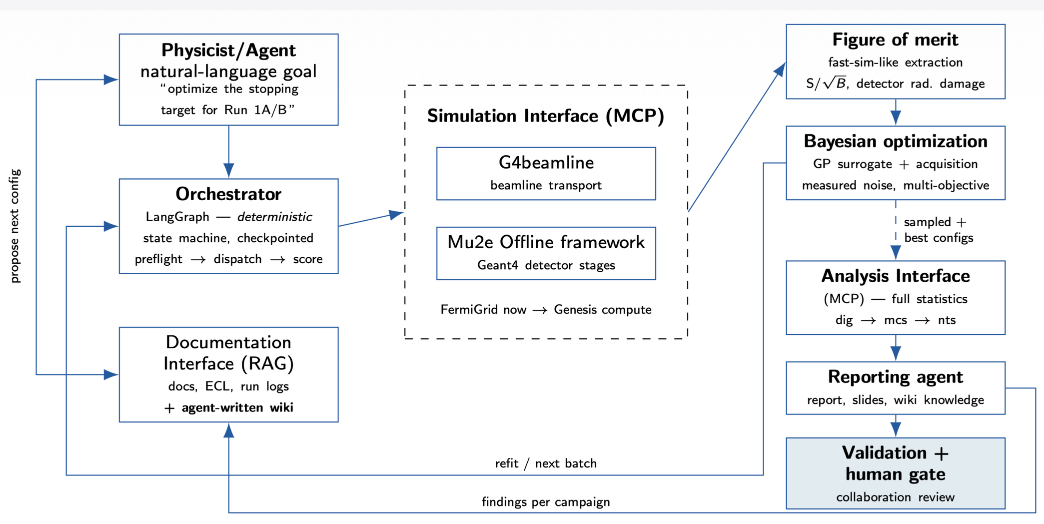
MCP client → Analysis and Knowledge-Base MCP servers → Mu2e computing, HAAI toolkit, Genesis infrastructure.

Workflow (i)

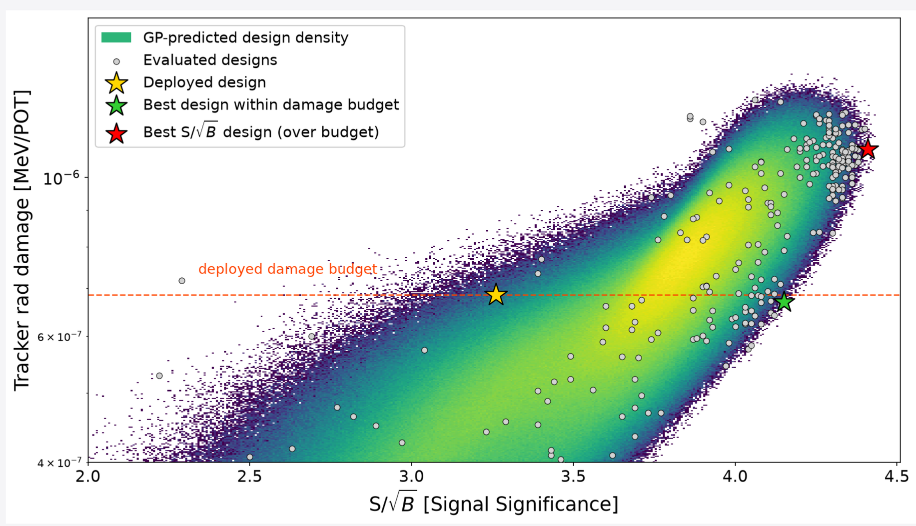
Agentic experiment optimization for Run-1

Challenge. A Run-1 configuration spans detector geometry, beamline settings and pipeline choices. Evaluating one scenario takes months of manual simulation and analysis work, and a new student needs up to a year to learn the chain.

Agentic execution. A physicist states the objective in natural language (e.g. "maximize sensitivity given a 10% reduction in solenoid field"). The agent assembles the configuration, launches Geant4 campaigns and analysis on FermiGrid and evaluates the returned figures of merit. Bayesian optimization handles the local search; the agent moves between qualitatively different configurations and diagnoses failed jobs.

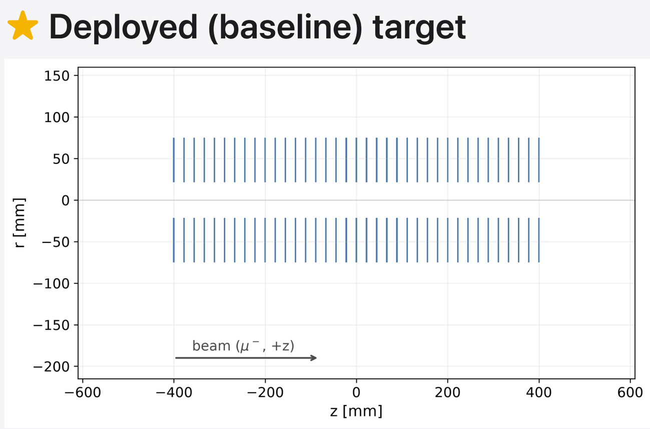


The loop runs end to end today: LangGraph orchestrator, Simulation and Analysis interfaces, GP surrogate with qNEHVI multi-objective acquisition (~150 CPU-h per evaluation), reporting agent, human gate. Campaigns write back to an optimization memory. Next: IPA, secondary stopping target, production target, cooling wedges.

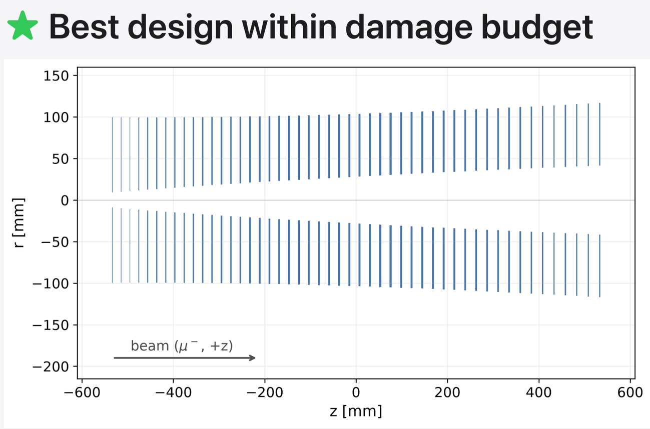


Proof of principle on the stopping target: GP-predicted design density, $S/\sqrt{B}$ vs. tracker radiation damage. The optimizer finds a design (green) beating the deployed one (yellow) at the same damage budget.

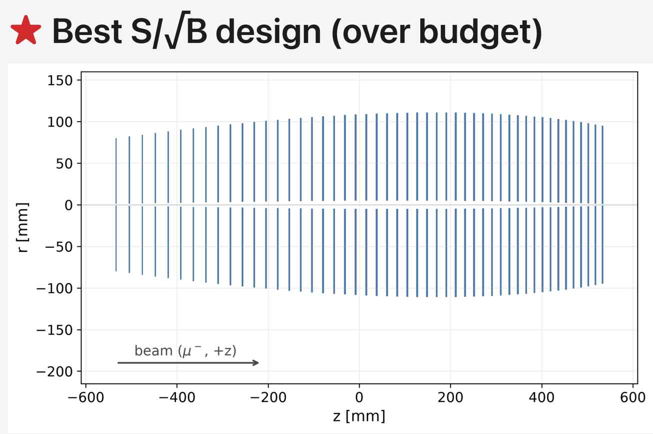
★ Deployed (baseline) target



★ Best design within damage budget



★ Best $S/\sqrt{B}$ design (over budget)



Impact. Iteration cycles drop from weeks to days. The collaboration can respond to changing constraints and explore configurations that would otherwise go untested, and expert tools become available to everyone, including new students.

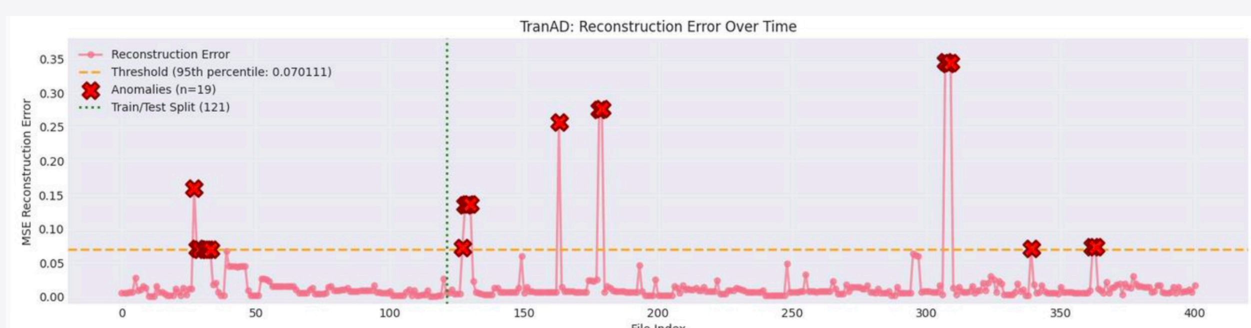
Workflow (ii)

AI-driven data quality monitoring for the Cosmic Ray Run

Challenge. Offline DQM plots are reviewed by shifters. Time allows only the most obvious anomalies to be investigated, and diagnosing a flagged issue means hours of searching logbooks, run databases and past incidents.

ML anomaly detection. HAAI transformer models (TranAD: unsupervised, learns temporal and multi-channel correlations) are integrated into the Offline DQM pipeline, validated with injected anomalies and benchmarked against non-ML methods on the Fermilab Elastic Analysis Facility GPUs.

Agentic DQM. After each run is processed, the agent compares plots against references, retrieves anomaly scores, then searches the electronic logbook, shifter logs, run-log and configuration databases for entries correlated in time and for past occurrences. The shifter receives a summary with links to the sources. Anything unresolved is escalated with suggested diagnostics.



Mu2e interface to the HAAI 9-model anomaly framework, running on the EAF GPU node. TranAD reconstruction error over CRV data files; a synthetic CRV fault is detected above the 95th-percentile threshold.

Impact. Every Cosmic Ray Run dataset is checked. Context assembly drops from hours to minutes, shifters spend their time on judgment rather than searching, and the same data allow a direct comparison with conventional DQM.