Brookhaven National Laboratory

Hunting for TREASURE in HEP collider data: Tokenized Representations for Energy-frontier AI Searches via Understanding and REasoning

Lead Lab	PI
Brookhaven National Laboratory	Viviana Cavaliere
Collaborating Labs	**Institutional PIs**
ANL	Walter Hopkins
FNAL	Kevin Pedro
LBNL	Paolo Calafiura
SLAC	Michael Kagan
DOE/SC Program Office:	High Energy Physics
DOE/SC Program Office Technical Contact:	Jeremy Love
Call	Office of High Energy Physics (HEP) American Science Cloud (AmSC) Intelligent Data Activities (IDAs) Pilots

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

U.S. DEPARTMENT OF ENERGY

1 Introduction and Scientific Case

The main barrier to AI-driven discovery in particle physics is that experiments generate exabytes of data in incompatible formats across different detectors. **TREASURE's core mission is transforming massive, heterogeneous HEP collider datasets into AI-ready, tokenized representations to unlock the full discovery potential of cross-experimental AI.**

Thanks to our extensive domain expertise in AI/ML for HEP, **TREASURE** will provide the ASCR Transformational AI Models Consortium (ModCon) [6] with curated, tokenized experimental physics datasets at unprecedented scales, as well as standards and tools to process the ever-growing data volumes from the current generation of collider experiments. Our comprehensive metadata schemas, quality assessment protocols, and community access frameworks will become templates for the AI Consortium to adapt for other scientific domains, multiplying the impact of our HEP-focused development effort across the broader Office of Science mission.

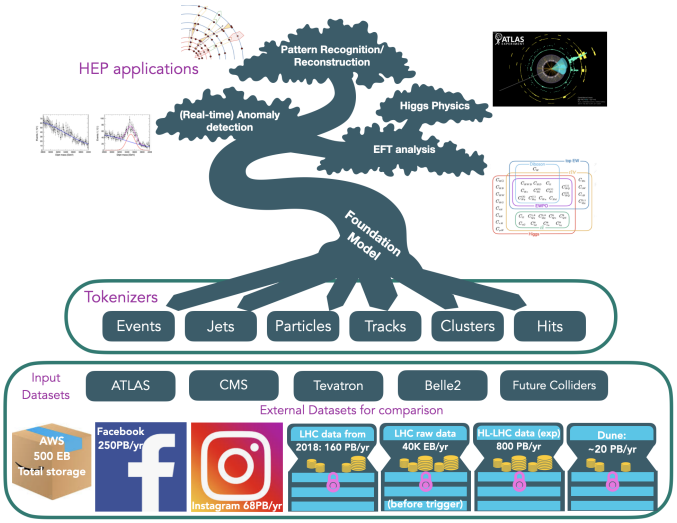

Figure 1: Sketch of the project, with external dataset sizes [1, 2] for comparison. Images from Refs. [3–5].

This work directly advances HEP's mission by: (1) preparing tokenized datasets and establishing community standards for AI-ready HEP collider data that will persist beyond this pilot; (2) demonstrating that properly prepared multi-modal datasets can enable discovery capabilities beyond single-experiment data; and (3) providing prototype foundation models fine-tuned for different physics tasks. Our vision is to create the largest curated, tokenized particle physics dataset ever assembled, drawn from the LHC and other past and future colliders. Moreover by enabling AI models to learn correlations across multiple experiments and collision energies, TREASURE will enhance sensitivity to fundamental parameters including Higgs couplings, electroweak precision observables, and signatures of physics beyond the Standard Model, ultimately improving constraints on theoretical models and guiding the search for new particles and interactions. Figure 1 shows a sketch of the project, starting from the input datasets we plan to use to the final physics tasks we plan to tackle.

2 Technical Approach and Methodology

2.1 AI-Ready Dataset Curation and Standardization

TREASURE's primary focus is delivering AI-ready HEP collider data produced by comprehensive data curation frameworks. We will develop these frameworks to transform raw HEP collider data into AI-ready formats using tokenization based on standard methods adapted to HEP. The frameworks will also include quality assessment strategies. Our approach addresses the full data preparation pipeline:

Cross-Experimental Simulation & Data Curation Framework: TREASURE will collaborate with the Transformational Model Consortium's Data Broker and Standards (ModCon DBS) [6] team to establish and adopt common AI data standards for scientific data. Our main goals include: (1) curating, providing, and transforming EB-scale datasets; (2) establishing a unified metadata schema that documents the format, content, and provenance of each dataset; (3) defining standard units and coordinate systems; and (4) developing a shared data model, initially based on widely adopted graph and point cloud data structures [7]. Our vision is to facilitate seamless

data integration from multiple experiments in different colliders. We will also combine recorded experimental data with simulation to help AI models learn robust physics representations rather than experiment-specific artifacts. In this context, we will recommend best practices for ground truth labeling. To clearly document the intent and capabilities of our datasets, we will accompany them with clear usage guidelines and one or more exemplary models that provide a starting point and benchmark for model development. Where appropriate, we will collaborate with the FAIR Universe project [8] and related activities in AmSC to organize machine learning challenges and leaderboards based on TREASURE datasets.

AI-Readiness Assessment, Documentation, & Enhancement: Each dataset will undergo systematic evaluation for AI-Readiness, including completeness analysis, bias detection, statistical characterization, and scalability assessment. In collaboration with the authors of AIDRIN [9], we will develop tools to evaluate the AI-Readiness of each dataset, with the goal of identifying and possibly correcting common issues that hinder effective AI training, such as duplicate or incomplete entries, inconsistent ground truth labeling, and ill-defined data semantics.

Multi-Level Tokenization Standards: Our tokenization strategy converts HEP collider data into standardized token sequences while preserving essential physics information. Our tokenizers will be based on tools currently in development within the community, such as Vector Quantized Variational Auto-Encoders (VQ-VAE) adapted to collider data, building on existing work from our team [10, 11]. Further, we will explore encoding approaches such as xVal for continuous numerical values [12]. Starting from ongoing efforts [13], we will combine low-level detector responses (hits, energy deposits), mid-level reconstructed objects (tracks, particles, jets), and high-level event features (topology, kinematics) to create AI-optimized multi-scale representations. Each level will include comprehensive metadata to document data provenance, processing steps, and quality metrics.

Compression of Tokenized Data: A critical component for scaling AI-ready HEP collider datasets is the development of intelligent compression techniques that dramatically reduce storage and bandwidth requirements while preserving the information essential for diverse particle physics tasks. We will implement compression schemes that exploit the multi-level structure of tokenized particle physics data, using variable-length encoding for common physics patterns. Our approach will also include learned compression for compact representations of detector responses that maintain maximum possible physics fidelity, as well as optimizations for real-time applications [14, 15]. These compression methods will be validated against physics benchmarks to ensure that compressed datasets enable the same discovery capabilities as uncompressed data while reducing storage requirements by orders of magnitude—a crucial capability for managing the exabyte-scale datasets anticipated from current and future HEP experiments.

2.2 Foundation Model Prototype and Assessment

Dataset preparation is our primary focus. To demonstrate the advantages of our data encoding approaches, we will train prototype HEP domain-specific foundation models to learn general physical principles from the encoded data, which can subsequently be used for physics tasks.

Model Architecture & Training: Our foundation models will be engineered to learn effectively from multiple input types (from low-level sensor hits to high-level event topologies) to understand the hierarchies encoded in the data. We will investigate supervised and self-supervised training approaches suitable for physics data with objectives including masked prediction [10] and contrastive representation learning [16–18], shown to improve domain adaptation and robustness [19]. To further enable optimal learning from tokenized data, we will explore model architectures that incorporate the physical principles that underlie our data, including detector geometry, particle trajectories, and conservation laws, building on recent successes [20].

Cross-Experiment Validation: We will validate our data preparation approach by training on tokenized data from one experiment and testing the model's ability to understand and reproduce physics consistent with another experiment. We will also train on data from multiple experiments

concurrently to demonstrate the enhanced predictive performance from this multi-modal learning. This will demonstrate that proper tokenization enables domain-independent learning of underlying physical principles and therefore provides cross-experimental capabilities.

Benchmark Dataset Applications: In addition to curated datasets, we will provide benchmark physics tasks: object and event classification, pattern recognition (clustering, tracking, particle flow), physics parameter extraction, and (offline and real-time) anomaly detection. Performance on these tasks will be used to establish standard evaluation metrics to assess different foundation models developed within TREASURE, with ASCR partners, and by the HEP community.

3 Timeline and Milestones

The collaborative work across participating laboratories is organized into multiple objectives and milestones spanning two years, as detailed in Table 1.

Year 1 Objectives

The first year focuses on establishing foundational infrastructure and standards. The data infrastructure development phase involves creating a cross-experiment data curation framework and common data model for LHC experiments, led by BNL, LBNL, SLAC, and FNAL. This includes establishing metadata standards for experimental data and trained models, implementing dataset management systems with selection and filtering capabilities, and conducting AI-readiness evaluations in collaboration with ASCR. Concurrently, teams will develop benchmarks and challenges for key physics tasks. BNL and LBNL will create data challenges for tracking, SLAC will focus on clustering and particle flow, while FNAL and SLAC will tackle jet reconstruction. Additionally, FNAL and BNL will establish anomaly detection benchmarks for real-time and offline analysis.

Establishing tokenization standards and preparing datasets are critical components. We will develop common protocols using current tools in development within the community,(VQ-VAE based tokenization) and each lab will focus on enabling protocols for specific data types: BNL for ATLAS tracking and triggers, ANL for DAQ and calorimeter systems, SLAC for calorimeter systems and jets, FNAL for CMS jets and triggers, and LBNL for tracks. Teams will create curated, tokenized datasets from LHC Open Data [21], complete with comprehensive documentation and usage examples. Validation metrics will ensure tokenization preserves essential physics information, supported by robust version control and community access protocols.

The year culminates with **Milestone 1**: the first release of our tokenized datasets from the LHC Open Data to the AI consortium and our data curation framework synchronized with the AmSC Model Consortium. We will also provide a prototype model for jet tagging—including Higgs boson decays—and event classification, to test the quality of the tokenized datasets.

Year 2 Objectives The second year emphasizes integration and application. The team will contribute to AmSC data standards development, that will be adopted for all TREASURE datasets. We will develop scalable dataset tokenization and foundation model training workflows in partnership with the ModCon Best Practices for Scientific Workflows and HEP-CCE Scalable ML teams.

The tokenization capabilities will expand to the full hierarchy of experimental data including low-level detector data (such as hits and clusters), incorporating smart compression, validation systems and multi-scale VQ-VAE techniques (BNL, ANL, SLAC). This will be particularly helpful for real-time applications. Having access to tokenized trigger data of reduced size would open the possibility of performing triggerless analysis by saving partial events before the trigger is applied. In **Milestone 2** we will release low-level detector data and trigger data for the Open LHC datasets and the updated tokenization schemes coordinated with the AmSC Consortium. A prototype model for tracking pattern recognition will be released as well for testing low-level detector data.

Multi-experiment integration is critical, and therefore we will expand the project to include: Tevatron data [22] (FNAL), which will complement the LHC with data (pp interactions) produced by proton-antiproton interactions; Belle II data (BNL), which will add e+e- collider data useful for future planned colliders; and future colliders' simulated datasets (BNL, LBNL, SLAC). SLAC,

FNAL, and BNL, in collaboration with their ASCR teams, will train a foundation model (FM) using these extensive datasets. Physics applications come to the forefront as teams fine-tune foundation models for different tasks, which will reflect data challenges defined in year 1. Each lab will work on a different application of the trained FM: Higgs rediscovery in LHC Open Data (BNL, SLAC) and Tevatron data (FNAL); cross-experiment multidimensional effective field theory fits (ANL); data reconstruction (SLAC) and data quality monitoring (ANL, LBNL, and BNL) using low-level detector data; cross-experiment and real-time anomaly detection by BNL and FNAL; and triggerless analysis applications (BNL). **Milestone 3** will be the release of the additional datasets listed above and the first cross-experiment foundation model with fine-tuned versions for different physics tasks.

	FY2026				FY2027			
Develop common framework and data model for LHC data	■	■	■	■				
Conduct dataset AI-readiness evaluation (with ASCR)			■	■	■	■		
Establish experiment-specific tokenization protocols	■	■	■	■				
Create curated, tokenized datasets from LHC Open Data	■	■	■	■				
Train Prototype foundation models on LHC dataset		■	■	■	■	■		
Milestone 1: Create curated, tokenized datasets from LHC Open Data and baseline model				■				
Scale dataset tokenization and foundation model training					■	■		
Extend tokenization to low-level detector data					■	■	■	■
Milestone 2: Release tokenization standards for low-level detector data and prototype model							■	
Incorporate additional datasets						■	■	■
Fine-tune foundation models for different tasks						■	■	■
Milestone 3: release additional datasets and the first cross-experiment foundation model								■

Table 1: Timetable for the proposed activities

4 Team Structure and Management

Our multi-laboratory team leverages unique capabilities across DOE facilities:

BNL (Lead lab): the BNL team has years long experience in machine learning pipelines for tracking and jet reconstruction in real-time applications [23]. They have started work towards tokenization of low-level detector data, compression and a cross-frontier foundation model to perform tracking that works for ATLAS and for Nuclear Physics experiments [13–15]. BNL provides world-class AI infrastructure and expertise in developing machine learning models and in data-management and scaling.

LBNL: The LBNL team has nearly ten years of experience in curating benchmark datasets and organizing ML challenges, such as TrackML [24], the Fair Universe challenge platform [8], and the 10^9 jets OmniLearned dataset [25]. LBNL developed cross-experimental MLpattern recognition frameworks [26–28] and foundation models for data analysis [29] with associated simulated datasets. Building on this experience, LBNL will contribute to the creation of a data curation framework, a data model for multi-modal HEP collider data, and tokenization schemes.

SLAC: SLAC has over a decade of experience in deep learning and AI for HEP [30–32], and continues to develop novel methods and applications for AI in fundamental physics [33–45] and to deliver high impact applications on ATLAS, e.g. in heavy flavour and boosted jet tagging [46–51]. Recent work established some of the first approaches for foundation models in HEP [10, 18, 52], including developing tokenizers for HEP collider data [10, 11]. SLAC has recently develop large scale ATLAS [53] and pheno [54–57] open datasets for AI application, and co-organized several

HEP collider data challenges [58–60].

FNAL: FNAL's experience with AI applications in HEP spans from the Tevatron [61] to the first evidence for boosted $H \to b\bar{b}$ [62–64] and tasks such as reconstruction [65] and background prediction [66]. FNAL has integrated industry approaches and tools [67–71] and created the Fast Machine Learning Lab [72] for real-time and on-detector applications [73–81]. Recently, FNAL has produced leading results in generative AI [82, 83], jet classification with transformers [84], and the first AI-ready open dataset of 160M jets from CMS [85]. FNAL's anomaly detection efforts include data challenges [86, 87], novel methods and results [88–90], and the first trigger-level AI [91].

ANL: Argonne has contributed to various aspects of the ATLAS trigger and data acquisition system, has studied compression schemes for the ATLAS physics analysis data format [92], is currently working on building an ML-based cross-experimental data quality monitoring (DQM) toolkit, and has worked to improve SMEFT interpretations [93, 94].

Management: The TREASURE activities will be managed via monthly coordination meetings, quarterly milestone reviews, and annual community workshops. Each institution leads specific deliverables, ensuring balanced contributions and efficient progress toward shared objectives.

5 Risk Assessment and Mitigation

TREASURE faces manageable technical and organizational risks. The primary technical risk is the loss of physics information during tokenization, which we mitigate through systematic validation using physics benchmarks and maintaining parallel analysis pipelines during development. Cross-experimental standardization challenges from differing detector technologies are addressed through our phased approach, starting with LHC Open Data before expanding to other experiments and lower-level data. Our focus on fundamental data standards ensures lasting value despite evolving AI architectures, while parallel development of established and novel approaches ensures timely delivery. The main organizational risk is that the collider community does not adopt the data standards and tools developed by TREASURE and, more broadly, by ModCon DBS. To mitigate this risk, we will leverage our direct connection with the LHC OpenData initiative and immediately start collaborating with them on defining data and metadata standards and prioritizing the developments needed by the community. We expect that AmSC will quickly make available abundant storage and computing resources. To mitigate the risk of delays in AmSC infrastructure availability, we will apply for allocations at NERSC and other resources like BNL's SDCC and SLAC's S3DF.

6 AmSC Integration, Expected Outcomes, and Community Impact

TREASURE will deliver critical data infrastructure enabling AI-driven discovery in particle physics and establish HEP as a leader in scientific data curation. Immediate outcomes include: standardized tokenization protocols for collider experiments, curated multi-experiment datasets with AI-readiness validation, dataset challenges with open source baseline models, and AmSC infrastructure integration. Success will be measured by: (1) AI-ready dataset delivery in tokenized and non-tokenized forms; (2) challenges, benchmarks, and baseline models driving dataset utilization; (3) demonstrations of multi-modal AI capabilities; (4) transformative cross-experimental discoveries; and (5) community adoption of standards impacting broader scientific computing.

TREASURE aligns with and bolsters the AmSC vision of shared scientific computing infrastructure. Our standardized, tokenized datasets will become foundational AmSC community resources, demonstrating how domain-specific data preparation can unlock AI capabilities across scientific disciplines. By collaborating with ModCon, our HEP collider data standards can inform and evolve with the broader SC data preparation methodologies. We will work with the FAIR universe team [8] and the AmSC project to develop and provide community access to datasets, models, and ML challenges. In this way, our AI-readiness evaluation protocols, metadata schemas, and community access frameworks can help establish best practices for the community, enable fast integration into AmSC, and influence how AmSC and the broader scientific computing community approach AI infrastructure development.

APPENDIX 1: BIBLIOGRAPHY & REFERENCES CITED

References

[1] W. Bhimji, D. Carder, E. Dart, J. Duarte, I. Fisk, R. Gardner, C. Guok, B. Jayatilaka, T. Lehman, M. Lin, C. Maltzahn, S. McKee, M. S. Neubauer, O. Rind, O. Shadura, N. V. Tran, P. van Gemmeren, G. Watts, B. A. Weaver, and F. Würthwein, *Snowmass 2021 Computational Frontier CompF4 Topical Group Report: Storage and Processing Resource Access*, 2022. arXiv:2209.08868 [physics.comp-ph].

[2] L. R. Luca Clissa, Mario Lassnig, *How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry*, 2023. Front. Big Data **6** (2023) 1271639.

[3] V. Belis, P. Odagiu, and T. K. Aarrestad, *Machine learning for anomaly detection in particle physics*, Rev. Phys. **12** (2024) 100091, arXiv:2312.14190 [physics.data-an].

[4] J. Ellis, M. Madigan, K. Mimasu, V. Sanz, and T. You, *Top, Higgs, Diboson and Electroweak Fit to the Standard Model Effective Field Theory*, JHEP **04** (2021) 279, arXiv:2012.02779 [hep-ph].

[5] A. Collaboration, *ATLAS Event Displays: Higgs boson decaying to two b-quarks*, https://cds.cern.ch/record/2636049.

[6] D. call. https://science.osti.gov/-/media/grants/pdf/lab-announcements/2025/LAB-25-3560-000001.pdf.

[7] S. Thais, P. Calafiura, G. Chachamis, G. DeZoort, J. Duarte, S. Ganguly, M. Kagan, D. Murnane, M. S. Neubauer, and K. Terao, *Graph Neural Networks in Particle Physics: Implementations, Innovations, and Challenges*, 2022. arXiv:2203.12852 [hep-ex].

[8] FAIR Universe Collaboration. https://fair-universe.lbl.gov/.

[9] K. Hiniduma, S. Byna, J. L. Bez, and R. Madduri, *AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI*, in *Proceedings of the 36th International Conference on Scientific and Statistical Database Management*, SSDBM 2024, p. 1–12. ACM, July, 2024.

[10] T. Golling, L. Heinrich, M. Kagan, S. Klein, M. Leigh, M. Osadchy, and J. A. Raine, *Masked particle modeling on sets: towards self-supervised high energy physics foundation models*, Mach. Learn. Sci. Tech. **5** (2024) 035074, arXiv:2401.13537 [hep-ph].

[11] M. Leigh, S. Klein, F. Charton, T. Golling, L. Heinrich, M. Kagan, I. Ochoa, and M. Osadchy, *Is Tokenization Needed for Masked Particle Modeling?*, Mach. Learn. Sci. Tech. **6** (2025) 025075, arXiv:2409.12589 [hep-ph].

[12] S. Golkar, M. Pettee, M. Eickenberg, A. Bietti, M. Cranmer, G. Krawezik, F. Lanusse, M. McCabe, R. Ohana, L. Parker, B. R.-S. Blancard, T. Tesileanu, K. Cho, and S. Ho, *xVal: A Continuous Number Encoding for Large Language Models*, in *NeurIPS 2023 AI for Science Workshop*. 2023. arXiv:2310.02989 [stat.ML]. https://openreview.net/forum?id=KHDMZtoF4i.

[13] D. Park, S. Li, Y. Huang, X. Luo, H. Yu, Y. Go, C. Pinkenburg, Y. Lin, S. Yoo, J. Osborn, J. Huang, and Y. Ren, *FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics*, 2025. arXiv:2508.14087 [cs.LG].

[14] Y. Huang, Y. Ren, S. Yoo, and J. Huang, *Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder*, 2021. arXiv:2111.05423 [cs.LG].

[15] Y. Huang, Y. Ren, S. Yoo, and J. Huang, *Fast 2D Bicephalous Convolutional Autoencoder for Compressing 3D Time Projection Chamber Data*, 2023. arXiv:2310.15026 [stat.ML].

[16] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, *Supervised contrastive learning*, in *Proceedings of the 34th International*

Conference on Neural Information Processing Systems, NIPS '20. Curran Associates Inc., Red Hook, NY, USA, 2020. arXiv:2004.11362 [cs.LG].

[17] A. Gandrakota, L. H. Zhang, A. Puli, K. Cranmer, J. Ngadiuba, R. Ranganath, and N. Tran, *Robust anomaly detection for particle physics using multi-background representation learning*, Mach. Learn. Sci. Tech. **5** (2024) 035082, arXiv:2401.08777 [hep-ex].

[18] P. Harris, J. Krupa, M. Kagan, B. Maier, and N. Woodward, *Resimulation-based self-supervised learning for pretraining physics foundation models*, Phys. Rev. D **111** (2025) 032010, arXiv:2403.07066 [hep-ph].

[19] A. Ćiprijanović, A. Lewis, K. Pedro, S. Madireddy, B. Nord, G. N. Perdue, and S. M. Wild, *DeepAstroUDA: semi-supervised universal domain adaptation for cross-survey galaxy morphology classification and anomaly detection*, Mach. Learn. Sci. Tech. **4** (2023) 025013, arXiv:2302.02005 [astro-ph.GA].

[20] H. Qu, C. Li, and S. Qian, *Particle Transformer for Jet Tagging*, in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, p. 18281. PMLR, 17–23 Jul, 2022. arXiv:2202.03772 [hep-ph]. https://proceedings.mlr.press/v162/qu22b.html.

[21] CERN, *LHC Open Data*, https://opendata.cern.ch.

[22] S. Amerio et al., *Data preservation at the Fermilab Tevatron*, Nucl. Instrum. Meth. A **851** (2017) 1, arXiv:1701.07773 [hep-ex].

[23] H. Abidi, A. Boveia, V. Cavaliere, D. Furletov, A. Gekow, C. W. Kalderon, and S. Yoo, *Charged Particle Tracking with Machine Learning on FPGAs*, 2022. arXiv:2212.02348 [physics.ins-det].

[24] S. Amrouche et al., *The Tracking Machine Learning challenge : Accuracy phase*, in *The NeurIPS '18 Competition: From Machine Learning to Intelligent Conversations*. 4, 2019. arXiv:1904.06778 [hep-ex].

[25] *The OmniLearned Data Repository*, https://omnilearned.nersc.gov/.

[26] The Exa.Trkx Project. https://exatrkx.github.io.

[27] The ATLAS GNN4ITk group. https://gitlab.cern.ch/gnn4itkteam/acorn.

[28] A. Aurisano, V. Hewes, G. Cerati, J. Kowalkowski, C. S. Lee, W. Liao, D. Grzenda, K. Gumpula, and X. Zhang, *Graph neural network for neutrino physics event reconstruction*, Phys. Rev. D **110** (2024) 032008, arXiv:2403.11872 [physics.data-an].

[29] J. Ho, B. R. Roberts, S. Han, and H. Wang, *Pretrained Event Classification Model for High Energy Physics Analysis*, 2024. arXiv:2412.10665 [hep-ph].

[30] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, JHEP **02** (2015) 118, arXiv:1407.5675 [hep-ph].

[31] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, *Jet-images — deep learning edition*, JHEP **07** (2016) 069, arXiv:1511.05190 [hep-ph].

[32] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, and T. Wongjirad, *Machine learning at the energy and intensity frontiers of particle physics*, Nature **560** (2018) 41.

[33] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/48ab2f9b45957ab574cf005eb8a76760-Paper.pdf.

[34] S. Shirobokov, V. Belavin, M. Kagan, A. Ustyuzhanin, and A. G. Baydin, *Black-Box Optimization with Local Generative Surrogates*, in *Advances in Neural Information Processing Systems*, vol. 33, p. 14650. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/a878dbebc902328b41dbf02aa87abb58-Paper.pdf.

[35] M. Vandegar, M. Kagan, A. Wehenkel, and G. Louppe, *Neural Empirical Bayes: Source Distribution Estimation and its Applications to Simulation-Based Inference*, in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, vol. 130 of *Proceedings of Machine Learning Research*, p. 2107. PMLR, 13–15 Apr, 2021. https://proceedings.mlr.press/v130/vandegar21a.html.

[36] S. Golkar, M. Kagan, and K. Cho, *Continual Learning via Neural Pruning*, 2019. arXiv:1903.04476 [cs.LG].

[37] L. Heinrich and M. Kagan, *Differentiable Matrix Elements with MadJax*, J. Phys. Conf. Ser. **2438** (2023) 012137, arXiv:2203.00057 [hep-ph].

[38] A. Kofler, V. Stimper, M. Mikhasenko, M. Kagan, and L. Heinrich, *Flow annealed importance sampling bootstrap meets differentiable particle physics*, Mach. Learn. Sci. Tech. **6** (2025) 025061, arXiv:2411.16234 [hep-ph].

[39] M. Aehle, M. Novák, V. Vassilev, N. R. Gauger, L. Heinrich, M. Kagan, and D. Lange, *Optimization using pathwise algorithmic derivatives of electromagnetic shower simulations*, Comput. Phys. Commun. **309** (2025) 109491, arXiv:2405.07944 [physics.comp-ph].

[40] R. E. C. Smith, I. Ochoa, R. Inácio, J. Shoemaker, and M. Kagan, *Differentiable vertex fitting for jet flavor tagging*, Phys. Rev. D **110** (2024) 052010, arXiv:2310.12804 [hep-ex].

[41] F. Mokhtar, J. Pata, D. Garcia, E. Wulff, M. Zhang, M. Kagan, and J. Duarte, *Fine-tuning machine-learned particle-flow reconstruction for new detector geometries in future colliders*, Phys. Rev. D **111** (2025) 092015, arXiv:2503.00131 [hep-ex].

[42] E. E. Khoda et al., *Ultra-low latency recurrent neural network inference on FPGAs for physics applications with hls4ml*, Mach. Learn. Sci. Tech. **4** (2023) 025004, arXiv:2207.00559 [cs.LG].

[43] E. M. Metodiev, B. Nachman, and J. Thaler, *Classification without labels: learning from mixed samples in high energy physics*, Journal of High Energy Physics **2017** (2017) no. 10, 174.

[44] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler, *OmniFold: A Method to Simultaneously Unfold All Observables*, Phys. Rev. Lett. **124** (2020) 182001, arXiv:1911.09107 [hep-ph].

[45] M. Paganini, L. de Oliveira, and B. Nachman, *CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, Phys. Rev. D **97** (2018) 014021, arXiv:1712.10321 [hep-ex].

[46] ATLAS Collaboration, G. Aad et al., *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*, Eur. Phys. J. C **83** (2023) 681, arXiv:2211.16345 [physics.data-an].

[47] A. Collaboration, *Transforming jet flavour tagging at ATLAS*, 2025. arXiv:2505.19689 [hep-ex].

[48] ATLAS Collaboration, *Transformer Neural Networks for Identifying Boosted Higgs Bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS*, tech. rep., CERN, Geneva, 2023. https://cds.cern.ch/record/2866601. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2023-021.

[49] ATLAS Collaboration, *DeXTer: Deep Sets based Neural Networks for Low-p_T $X \to b\bar{b}$ Identification in ATLAS*, tech. rep., CERN, Geneva, 2022. https://cds.cern.ch/record/2825434.

[50] ATLAS Collaboration, *Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS*, tech. rep., CERN, Geneva, 2020. https://cds.cern.ch/record/2718948.

[51] ATLAS Collaboration, *Identification of Jets Containing b-Hadrons with Recurrent Neural*

Networks at the ATLAS Experiment, tech. rep., CERN, Geneva, 2017. https://cds.cern.ch/record/2255226.

[52] V. Mikuni and B. Nachman, *Solving Key Challenges in Collider Physics with Foundation Models*, 2025. arXiv:2404.16091 [hep-ph].

[53] ATLAS Collaboration, *ATLAS $t\bar{t}$ simulation for ML-based jet flavour tagging (JetSet)*, CERN Open Data Portal (2025).

[54] P. C. Harris, M. Kagan, J. Krupa, B. Maier, and N. Woodward, *RS3L: A jet tagging dataset for self-supervised learning based on re-simulation*, Mar., 2024. https://doi.org/10.5281/zenodo.10633815.

[55] I. Ochoa, R. Smith, L. Pereira Sánchez, and M. Kagan, *Dataset for flavour tagging R&D*, Aug., 2024. https://doi.org/10.5281/zenodo.13350327.

[56] F. Mokhtar, J. Pata, M. Kagan, D. Garcia, E. G. T. Wulff, M. Zhang, and J. M. Duarte, *Simulated datasets for detector and particle flow reconstruction: CLD detector model for FCC-ee, machine learning format*, Feb., 2025. https://doi.org/10.5281/zenodo.14930610.

[57] J. Pata, F. Mokhtar, M. Zhang, E. Wulff, D. Garcia, M. Kagan, and J. Duarte, *Simulated datasets for detector and particle flow reconstruction: CLIC detector, machine learning format*, Mar., 2025. https://doi.org/10.5281/zenodo.15062717.

[58] O. Amram et al., *CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation*, 10, 2024. arXiv:2410.21611 [physics.ins-det].

[59] G. Kasieczka et al., *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*, Rept. Prog. Phys. **84** (2021) 124201, arXiv:2101.08320 [hep-ph].

[60] W. Bhimji, P. Calafiura, R. Chakkappai, P.-W. Chang, Y.-T. Chou, S. Diefenbacher, J. Dudley, S. Farrell, A. Ghosh, I. Guyon, C. Harris, S.-C. Hsu, E. E. Khoda, R. Lyscar, A. Michon, B. Nachman, P. Nugent, M. Reymond, D. Rousseau, B. Sluijter, B. Thorne, I. Ullah, and Y. Zhang, *FAIR Universe HiggsML Uncertainty Challenge Competition*, 2024. arXiv:2410.02867 [hep-ph], https://arxiv.org/abs/2410.02867.

[61] B. H. Denby, *Neural Networks and Cellular Automata in Experimental High-energy Physics*, Comput. Phys. Commun. **49** (1988) 429–448.

[62] CMS Collaboration, *Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector*, CMS Detector Performance Summary CMS-DP-2018-046, CERN, 2018. http://cds.cern.ch/record/2630438.

[63] CMS Collaboration, *Performance of the mass-decorrelated DeepDoubleX classifier for double-b and double-c large-radius jets with the CMS detector*, CMS Detector Performance Summary CMS-DP-2022-041, CERN, 2022. http://cds.cern.ch/record/2839736.

[64] CMS Collaboration, A. M. Sirunyan et al., *Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **12** (2020) 085, arXiv:2006.13251 [hep-ex].

[65] Exa.TrkX Collaboration, X. Ju et al., *Graph Neural Networks for Particle Reconstruction in High Energy Physics detectors*, in *33rd Annual Conference on Neural Information Processing Systems*. 3, 2020. arXiv:2003.11603 [physics.ins-det].

[66] CMS Collaboration, A. Hayrapetyan et al., *Machine learning method for enforcing variable independence in background estimation with LHC data: ABCDisCoTEC*, arXiv:2506.08826 [hep-ex]. submitted to *Mach. Learn. Sci. Tech.*

[67] J. Duarte et al., *FPGA-accelerated machine learning inference as a service for particle physics computing*, Comput. Softw. Big Sci. **3** (2019) 13, arXiv:1904.08986 [physics.data-an].

[68] J. Krupa et al., *GPU coprocessors as a service for deep learning inference in high energy*

physics, Mach. Learn. Sci. Tech. **2** (2021) 035005, arXiv:2007.10359 [physics.comp-ph].

[69] D. S. Rankin et al., *FPGAs-as-a-Service Toolkit (FaaST)*, Oct., 2020. arXiv:2010.08556 [physics.comp-ph].

[70] CMS Collaboration, A. Hayrapetyan et al., *Portable Acceleration of CMS Computing Workflows with Coprocessors as a Service*, Comput. Softw. Big Sci. **8** (2024) 17, arXiv:2402.15366 [physics.ins-det].

[71] H. Zhao et al., *Track reconstruction as a service for collider physics*, JINST **20** (2025) P06002, arXiv:2501.05520 [physics.ins-det].

[72] *Fast Machine Learning Lab*, https://fastmachinelearning.org.

[73] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, JINST **13** (2018) P07027, arXiv:1804.06913 [physics.ins-det].

[74] V. Loncar et al., *Compressing deep neural networks on FPGAs to binary and ternary precision with HLS4ML*, Mach. Learn. Sci. Tech. **2** (2021) 015001, arXiv:2003.06308 [cs.LG].

[75] Y. Iiyama et al., *Distance-Weighted Graph Neural Networks on FPGAs for Real-Time Particle Reconstruction in High Energy Physics*, Front. Big Data **3** (2020) 598927, arXiv:2008.03601 [physics.ins-det].

[76] A. Heintz et al., *Accelerated Charged Particle Tracking with Graph Neural Networks on FPGAs*, in *34th Conference on Neural Information Processing Systems*. 11, 2020. arXiv:2012.01563 [physics.ins-det].

[77] T. Aarrestad et al., *Fast convolutional neural networks on FPGAs with hls4ml*, Mach. Learn. Sci. Tech. **2** (2021) 045015, arXiv:2101.05108 [cs.LG].

[78] F. Fahim et al., *hls4ml: An Open-Source Codesign Workflow to Empower Scientific Low-Power Machine Learning Devices*, in *tinyML Research Symposium 2021*. 3, 2021. arXiv:2103.05579 [cs.LG].

[79] G. Di Guglielmo et al., *A Reconfigurable Neural Network ASIC for Detector Front-End Data Compression at the HL-LHC*, IEEE Trans. Nucl. Sci. **68** (2021) 2179, arXiv:2105.01683 [physics.ins-det].

[80] J. Yoo et al., *Smart pixel sensors: towards on-sensor filtering of pixel clusters with deep learning*, Mach. Learn. Sci. Tech. **5** (2024) 035047, arXiv:2310.02474 [physics.ins-det].

[81] J. Dickinson et al., *Smartpixels: Towards on-sensor inference of charged particle track parameters and uncertainties*, arXiv:2312.11676 [hep-ex].

[82] O. Amram and K. Pedro, *Denoising diffusion models with geometry adaptation for high fidelity calorimeter simulation*, Phys. Rev. D **108** (2023) 072014, arXiv:2308.03876 [physics.ins-det].

[83] O. Amram and M. Szewc, *Data-Driven High-Dimensional Statistical Inference with Generative Models*, arXiv:2506.06438 [hep-ph].

[84] CMS Collaboration, *Search for highly energetic double Higgs boson production in the two bottom quark and two vector boson all-hadronic final state*, CMS Physics Analysis Summary CMS-HIG-23-012, CERN, 2024. http://cds.cern.ch/record/2904879.

[85] O. Amram, L. Anzalone, J. Birk, D. A. Faroughy, A. Hallin, G. Kasieczka, M. Krämer, I. Pang, H. Reyes-Gonzalez, and D. Shih, *Aspen Open Jets: unlocking LHC data for foundation models in particle physics*, Mach. Learn. Sci. Tech. **6** (2025) 030601, arXiv:2412.10504 [hep-ph].

[86] E. Govorkova, E. Puljak, T. Aarrestad, M. Pierini, K. A. Woźniak, and J. Ngadiuba, *LHC physics dataset for unsupervised New Physics detection at 40 MHz*, Sci. Data **9** (2022) 118, arXiv:2107.02157 [physics.data-an].

[87] T. Aarrestad et al., *The Dark Machines Anomaly Score Challenge: Benchmark Data and*

Model Independent Event Classification for the Large Hadron Collider, SciPost Phys. **12** (2022) 043, arXiv:2105.14027 [hep-ph].

[88] CMS Collaboration, *Wasserstein normalized autoencoder*, CMS Physics Analysis Summary CMS-PAS-MLG-24-002, CERN, 2024. http://cds.cern.ch/record/2911111.

[89] CMS Collaboration, *Machine-learning techniques for model-independent searches in dijet final states*, CMS Physics Analysis Summary CMS-MLG-23-002, CERN, 2025. http://cds.cern.ch/record/2938054.

[90] CMS Collaboration, V. Chekhovsky et al., *Model-agnostic search for dijet resonances with anomalous jet substructure in proton–proton collisions at* $\sqrt{s} = 13$ *TeV*, Rept. Prog. Phys. **88** (2025) 067802, arXiv:2412.03747 [hep-ex].

[91] CMS Collaboration, A. Gandrakota, *Realtime Anomaly Detection at the L1 Trigger of CMS Experiment*, PoS **ICHEP2024** (2025) 1025, arXiv:2411.19506 [hep-ex].

[92] C. Marcon, A. S. Mete, P. Van Gemmeren, and L. Carminati, *Optimizing ATLAS data storage: The impact of compression algorithms on ATLAS physics analysis data formats*, EPJ Web of Conf **295** (2024) 03027.

[93] B. Kriesten and T. J. Hobbs, *Anomalous electroweak physics unraveled via evidential deep learning*, Eur. Phys. J. C **85** (2025) 883, arXiv:2412.16286 [hep-ph].

[94] J. Gao, M. Gao, T. J. Hobbs, D. Liu, and X. Shen, *Simultaneous CTEQ-TEA extraction of PDFs and SMEFT parameters from jet and* $t\bar{t}$ *data*, JHEP **05** (2023) 003, arXiv:2211.01094 [hep-ph].

[95] M. Barisits et al., *Rucio - Scientific data management*, Comput. Softw. Big Sci. **3** (2019) 11, arXiv:1902.09857 [cs.DC].

APPENDIX 2: Data Management and Sharing 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 High Energy Physics (HEP), along with associated models and benchmarks, and make them widely accessible to the AmSC, HEP, 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. During this pilot project, we will focus on publicly available input data. All TREASURE curated datasets, along with associated models and other software tools, will be made publicly available.

Software: AI/ML models, data curation and management tools, and any other software associated with TREASURE datasets will be released using standard open source licenses, such as BSD 3-Clause License, GNU Public License, or MIT License. The exact license type will be determined once we identify the software dependencies for our work (if any) as well as the BNL and DOE policies and preferences, as they may affect which license we can use.

Software will be version-controlled and accessible via repositories like GitHub or GitLab. The versions of the software used for any publications will be documented, using either a git commit identifier string or a git tag. The landing page of the repository will also serve as the project website.

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 through DOE Data ID Service and/or Zenodo so that each release is guaranteed retrievable with a DOI. We will ensure that datasets produced are registered and have bidirectional links to other platforms used by the HEP community, such as CERN OpenData and HEPData. We will make our larger datasets available through distributed data management software deployed at DOE facilities and in the HEP community, like Rucio [95] and Globus.

TREASURE's 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, TREASURE datasets will be stored using the extensive data storage resources available at the participating laboratories.

In accordance with our lead laboratory's research and development standards, we will preserve all other digital data products using GitHub for code and other laboratory data storage, such as SharePoint, to apply appropriate management methods to the project's electronic information. 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.

APPENDIX 3: Budget narrative

The TREASURE project will support a comprehensive multi-laboratory effort focused on creating AI-ready HEP datasets. The budget is distributed across five DOE national laboratories, each contributing unique expertise and capabilities essential for project success. Personnel costs represent the primary budget component, reflecting the intensive data curation, tokenization development, and validation work required to transform heterogeneous HEP data into standardized, AI-ready formats. The budget is summarized in Table 2.

Table 2: Requested budget in US dollars by Institution and by Fiscal Year

Institution	FY26	FY27
BNL	$753k	$646k
ANL	$288k	$301k
LBNL	$360k	$240k
SLAC	$298k	$308k
FNAL	$346k	$254k
Total	$2.045M	$1.749M

6.1 Brookhaven National Laboratory (Lead Institution)

As project lead, BNL receives the largest allocation to support comprehensive coordination. BNL's budget supports specialized work on low-level data tokenization (for uses such as pattern recognition in tracking), trigger-level data tokenization (for uses for triggerless workflows and real-time anomaly detection) and software infrastructure, and physics applications relating to Higgs physics and anomaly detection.

Personnel:

- **1.5 FTE Postdocs :** Four specialized postdoctoral researchers: (1) 0.5 FTE postdoc (in the process of being hired) focusing on trigger-level data tokenization and cross-experiment anomaly detection, (2) 0.3 FTE postdoc helping with tokenization and inclusion of future collider datasets, (3) 0.2 FTE postdoc helping with the inclusion of the Belle 2 dataset, (4) 0.5 FTE postdoc specializing in intelligent data compression for real-time applications and foundation model deployment. Potential Postdocs at BNL to fill these roles include Iza veliscek to help with future collider data inclusion, Manfredi for the Belle II data inclusion, Ang Li for the tokenization efforts and Malige for the real-time applications.
- **0.4 FTE Computing Professional (Shuwei Ye):** Dedicated support for scaling foundation model training infrastructure and managing computational resources across participating institutions.
- **0.4 FTE Senior Scientists (Viviana Cavaliere):** 0.2 FTE senior scientist for overall project management and coordination, plus 0.2 FTE senior scientist for direct supervision of postdoctoral researchers and coordination with Nuclear Physics program at BNL.

Supporting Team Members:

- **Elizabeth Brost** (BNL, PO): liaise with the Nuclear Physics program and Intensity Frontier efforts of tokenization.
- **Haider Abidi** (BNL, PO): contribute to the definition of low-level detector data and help with real-time applications.
- **Yihui Ren** (BNL, Computer and Data Science department): help with training the Foundation Model on different datasets and new compression schemes for the low-level detector and trigger datasets.

Travel: BNL requests support in the form of travel for student exchanges between laboratories to facilitate collaborative training and knowledge transfer and for a small workshop.

6.2 Argonne National Laboratory

ANL's budget supports specialized work on low-level data tokenization (for uses such as data quality monitoring), software infrastructure, and physics applications relating to Standard Model Effective Field Theory interpretations:

Personnel (100% of budget):

- **1.0 FTE Postdoc:** Full-time postdoctoral researcher developing tokenization schemes for data acquisition and detector data quality monitoring applications.
- **0.1 FTE Staff Scientist:** Expert for physics applications.
- **0.2 FTE Computational Scientist:** Expert for ATLAS software infrastructure development and tokenization framework integration.

Supporting Team Members:

- **Azton Wells and Nesar Ramachandra** (ANL, Computational Science Division): with experience in developing foundation models for HEP, they will give feedback on tokenization schemes.
- **Walter Hopkins** (ANL, HEP): liaise with other TREASURE institutions, the Scaling ML group within the High-Energy Physics Center for Computing Excellence project, and data quality monitoring experts.

6.3 Lawrence Berkeley National Laboratory

LBNL will prioritize developing data models, infrastructure, and curation frameworks. According to the project timetable (Table 1), LBNL will focus its efforts during the initial 18 months of the project, with smaller contributions during the last six months towards achieving Milestone 3.

Personnel (100% of budget):

- **Beojan Stanislaus** (Physics Division PostDoc, 0.8 FTEs Year 1, 0.6 FTEs Year 2): Primary researcher for data curation methodologies and tokenization infrastructure development.
- **Julien Esseiva** (Scientific Data Division Software Engineer, 0.3 FTEs Year 1, 0.1 FTEs Year 2): Distributed data management and workflow development for data curation infrastructure and community access tools on grid, cloud, and HPC resources.
- **Graduate Student Research Assistant** (0.25 FTEs): Prototype models for validation of curation and tokenization infrastructure.

Supporting Team Members:

- **Jean Luca Bez and Surendra Byna** (LBNL SDD): as senior members of the AIDRIN team, they will contribute to the definition of HEP AI-Readiness standards and adapt the tools to check them.
- **Wahid Bhimji** (LBNL NERSC): liaise with the FAIR Universe project and with AmSC Infrastructure teams.
- **Paolo Calafiura** (LBNL SDD): liaise with other TREASURE institutions, the HEP-CCE project, and the US ATLAS program.
- **Zachary Marshall** (LBNL PD): liaise with the ATLAS and LHC Open Data initiatives.
- **Haichen Wang** (LBNL PD): contribute to the definition of data and metadata curation standards, consult on tokenization schemes and their Physics applications.

6.4 SLAC National Laboratory

SLAC will focus on helping develop the data curation, especially the data model and data / metadata standards, building tokenization models (that build on our prior work) with a focus on calorimeter and jet data and helping develop more advanced tokenizers, training domain-specific prototype foundation models on single and multi-experiment data (that build on our prior work),

using the model to provide baseline models for data challenges and benchmarks and to examine physics tasks like reconstruction and Higgs analysis.

Personnel (100% of Budget):

- **1.5 FTE Postdocs:** One postdoc will work on *year 1:* data curation, tokenizers, benchmarks (data prep), and foundation model training, and *year 2:* new tokenizer R&D foundation model training, and physics tasks. The 0.5 FTE postdoc will work on *year 1:* benchmarks (model and software preparation) and foundation model training, and *year 2:* benchmarks development on new low-level data available in year 2, foundation model fine-tuning for benchmarks, and physics tasks. Potential Postdocs at SLAC to fill these roles include Jeffrey Krupa, a AI/ML in HEP postdoc in the Energy Frontier group, and Samuel Klein, a AI/ML postdoc in the SLAC Machine Learning Initiative who did his PhD on AI in HEP projects.
- **0.05 FTE Senior Investigator:** Michael Kagan will serve as the EF lead PI who will oversee the execution of the SLAC team's work and provide team guidance, and provide feedback to the broader multi-lab effort.

6.5 Fermi National Accelerator Laboratory

FNAL's budget supports CMS-focused tokenization and physics applications:

Personnel (100% of budget):

- **0.25 FTE Associate Scientist (Abhijith Gandrakota**, FNAL Wilson Fellow): Intellectual leadership for the data deliverables, AI R&D tasks, and physics goals; supervision of other personnel.
- **0.50 FTE Research Associate**: Develop tokenization schemes using representation learning, along with metrics to assess the performance of the schemes on benchmark tasks. Curate the data to be delivered for use in the American Science Cloud, focusing initially on CMS jet and trigger data, and later adding Tevatron data from the CDF and D0 experiments. The resulting foundation model prototypes trained on this data will be employed for physics tasks including cross-experiment anomaly detection and Higgs rediscovery. Potential existing postdoctoral effort is available from Oz Amram [58, 82, 83, 85, 89, 90], Raghav Kansal [84], and new hire Sitian Qian [20].
- **0.14–0.40 FTE Computing Professional**: Establish and operate data movement and processing infrastructure to deliver the datasets, focusing first on CMS and then extending to the Tevatron. This process will include testing and providing feedback on the suitability of the common data format and processing framework to be established as part of the proposal. Potential existing effort is available from Scarlet Norberg (Software Developer).

Supporting Team Members:

- **Kevin Pedro** (FNAL Scientist): Oversight and guidance as FNAL PI, liaise with other TREASURE institutions.
- **Bo Jayatilaka** (FNAL Collider Physics Division Head): Expertise on Tevatron data as CDF SM convener and leader of the CDF data preservation project.