10 Million Particle Events: Enabling Foundation Models for Sparse 3D Inverse Problems

Omar Alterkait[1,4], **Sam Young**[2,3], **Ka Vang Tsang**[2], **Junjie Xia**[2],
Carolyn H. Smith[3], **Taritree Wongjirad**[1,4], **Kazuhiro Terao**[2]

[1]Tufts University [2]SLAC National Accelerator Laboratory [3]Stanford University
[4]The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Correspondence: {omar.alterkait@tufts.edu, kterao@slac.stanford.edu}

Abstract

Next-generation particle physics experiments require unprecedented machine learning capabilities to achieve their science goals. We propose generating 10 million particle detector events, the first dataset providing raw sensor waveforms paired with 3D ground truth at scale, enabled by GPU-accelerated JAX simulations achieving two orders of magnitude speedup over traditional CPU-based tools. This dataset will enable large-scale self-supervised training of foundation models for complex inverse problems in the particle physics and beyond.

1 AI Task Definition

Next-generation neutrino experiments such as the Deep Underground Neutrino Experiment (DUNE) [1] require machine learning models capable of achieving percent-level measurement precision on complex physics processes. The fundamental challenge involves a multi-level inverse problem that progresses through several stages of increasing complexity.

The forward process begins when particles interact in the detector, depositing energy and leaving trails of ionization electrons as they traverse the medium. In these detectors [4], ionization patterns drift through an electric field and detected using sensor arrays, recording projected 2D images where one spatial dimension becomes encoded in arrival time. This creates complex tomographic projections where 3D spatial information gets compressed into 2D sensor readings. Simultaneously, optical photons produced from the same energy depositions are collected using photosensors, providing complementary temporal and spatial information through a different physical mechanism.

The machine learning challenge requires inverting this entire process: starting from these 2D sensor projections and light signals, we must reconstruct the original 3D view of particle trajectories and extract the underlying physics. This reconstruction spans multiple scales, from sub-millimeter track features to meter-scale event topologies. The multi-modal nature adds complexity, as charge and light signals must be combined despite having different temporal resolutions and noise characteristics. Calibration of detector physics models represents another critical challenge where models must learn detector response patterns directly from data, accounting for variations in electronics and detector conditions.

Foundation models [5] trained through self-supervised learning on this data would learn robust representations of detector physics that can be adapted through transfer learning to specific experimental configurations. Such models could demonstrate scaling laws in scientific domains, showing how performance improves with data quantity and model size. We will create specialty datasets targeting specific physics challenges in collaboration with domain experts, ensuring these address real experimental needs such as challenging event topologies or particular background conditions.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

2 Dataset Rationale

Current machine learning approaches in neutrino physics face a fundamental bottleneck: existing simulations run on CPU farms, taking approximately 50 seconds per event, making large-scale dataset generation computationally prohibitive. Our complete simulation rewrite in JAX [3] leverages GPU parallelization and optimized memory-compute tradeoffs, reducing computation from 50 seconds on CPU to sub-0.5 seconds per event on A100 GPUs. This 100× acceleration through vectorized physics calculations transforms previously intractable dataset generation into achievable goals.

The dataset will contain 10 million particle events with randomly sampled particle types and kinematics to ensure unbiased coverage of the physics phase space. Each event includes complete ground truth hierarchy from initial particles through energy depositions to final sensor responses. The dual representation provides both the true 3D point clouds of ionization with sub-millimeter resolution spanning meter-scale volumes, and the corresponding 2D projections as waveform arrays with approximately 1500-2000 wire channels across time (see Appendix A for the complete data pipeline and visualization). Time-series light sensor data provides the complementary modality, capturing nanosecond-scale timing information across distributed photosensors. Rich labels include particle types, energies, interaction vertices, and detailed physics processes at each stage.

Using sparse data structures, each event requires approximately 20 MB of storage, totaling 200 TB for the complete dataset. At 10 million events, we exceed prior datasets like PILArNet [2] (300k events) by 30× while uniquely providing raw waveforms and optical signals that enable self-supervised learning approaches previously limited by data availability. We will generate this data on the NERSC supercomputing cluster, which has sufficient GPU resources for this scale of simulation.

3 Acceleration Potential

The immediate impact centers on enabling foundation models for particle physics, where a single model trained on this comprehensive dataset can be fine-tuned for specific experimental configurations, preventing the current redundant development where each experiment builds machine learning solutions from scratch. Appendix B details specific data challenges including tomographic reconstruction, 3D pattern recognition, and foundation model development that will drive algorithm advancement. This efficiency gain becomes critical as experiments prepare for data-taking, allowing teams to focus on experiment-specific optimizations rather than rebuilding basic capabilities.

This dataset provides an opportunity to advance spatially sparse 3D architectures, addressing the challenge of data that is globally sparse but locally information-dense. The inverse problem techniques developed here, reconstructing 3D structures from complex 2D projections, represent fundamental challenges that appear across scientific computing. The multi-modal aspects, where different sensor types capture complementary physics information with varying characteristics, push the boundaries of how neural networks can integrate heterogeneous data streams.

Beyond particle physics, this dataset provides a testbed for techniques applicable to sparse 3D reconstruction challenges across scientific domains. The core problem of tomographic reconstruction from incomplete data appears throughout medical imaging, where similar inverse problems require recovering 3D structures from limited projections. Architectures and optimization strategies developed for our extremely sparse data (under 2% active pixels with locally dense information) address challenges similar to those in autonomous driving where LiDAR produces inherently sparse 3D point clouds. While the underlying physics differs across applications, the 10 million event scale enables rigorous testing of whether neural architectures and optimization techniques generalize, providing empirical evidence about transferability that smaller datasets may not easily provide.

The infrastructure surrounding the dataset ensures lasting impact: open-source JAX simulations will be released alongside the data, enabling researchers to generate custom variants or extend to new detector configurations. A public data portal will host the datasets, trained models, and evaluation metrics, lowering barriers to entry for new researchers. Regular Data Olympics competitions will drive continuous algorithm development, providing structured challenges that advance the state of the art while building community engagement. Success will be measured through concrete metrics: establishment of state-of-the-art baseline models that future work must exceed, and documented adoption by particle physics experiments for their machine learning pipelines.

References

[1] B. Abi et al. Deep underground neutrino experiment (dune), far detector technical design report, volume i: Introduction to dune, 2020.

[2] Corey Adams et al. Pilarnet: Public dataset for particle imaging liquid argon detectors in high energy physics, 2020.

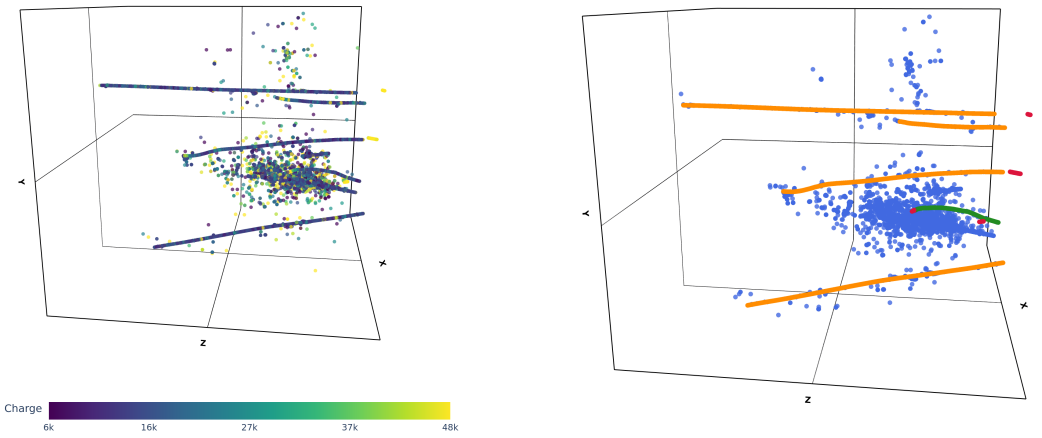
[3] James Bradbury et al. JAX: composable transformations of Python+NumPy programs, 2018.

[4] Krishanu Majumdar and Konstantinos Mavrokoridis. Review of liquid argon detector technologies in the neutrino sector. *Applied Sciences*, 11(6):2455, March 2021.

[5] Ce Zhou et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. 2023.

A Dataset Structure and Visualization

Our dataset captures the complete simulation pipeline from initial particle interactions through detector physics to final sensor readouts. The unique challenge stems from the detector's dual-readout geometry: particles drift toward **two independent readout planes** separated by a central cathode, with each side containing three wire planes at different orientations. This creates six distinct 2D projections of each 3D event, a richer but more complex reconstruction problem than typical tomographic imaging



(a) (b)

Figure 1: **3D Ground Truth.** Left: Ionization charge density. Right: Particle type labels (red: proton, orange: muon, green: pion, blue: electron).

Figure 1 shows the 3D ground truth that models must reconstruct from 2D projections. The data consists of 0.1 mm segments spanning the full 4.32 m³ detector volume, with each segment containing rich physics information: unique segment identifiers, parent particle IDs, interaction types, and complete particle ancestry trees. This hierarchical labeling enables models to learn not just particle classification but also the causal physics relationships between primary interactions and secondary particle production. The extreme density variations, from isolated tracks to dense overlapping cascades, require architectures that can handle sparse global structure with locally complex features.

The transformation from 3D truth to 2D projections involves multiple physics processes. Each readout wire extends across the detector plane and produces a signal whenever charge drifts past any point along its length. This collapses one spatial dimension entirely: a wire cannot distinguish where along its length the charge passed, only when it arrived. The three wire planes on each side are oriented at different angles to provide complementary projections that together enable 3D reconstruction, though the coverage remains incomplete compared to full tomographic imaging. During drift, electron clouds undergo diffusion, transforming sharp energy deposits into broader distributions that vary with drift distance.

Figure 2 shows the charge distribution after drift and diffusion effects. The electron clouds have spread from their original positions, with the amount of diffusion depending on the drift distance. The detector's central cathode splits the volume in half, with particles drifting either east or west depending on their position. Each side's three wire planes capture only the charge that drifts to that side, meaning the six projections come from two separate drift volumes rather than being complete views of the same space. This intermediate stage represents the charge as it arrives at the wire planes before any response effects are applied.

Figure 3 shows what the detector actually records after the complete response simulation. The two induction planes (U, V) record induced signals as charge drifts past, while the collection planes (Y) directly collect the arriving charge. This stage incorporates the complex detector response, including field variations and electronics shaping. While noise sources are included in the full simulation, they are omitted from this visualization for clarity. This is the raw data that any reconstruction algorithm

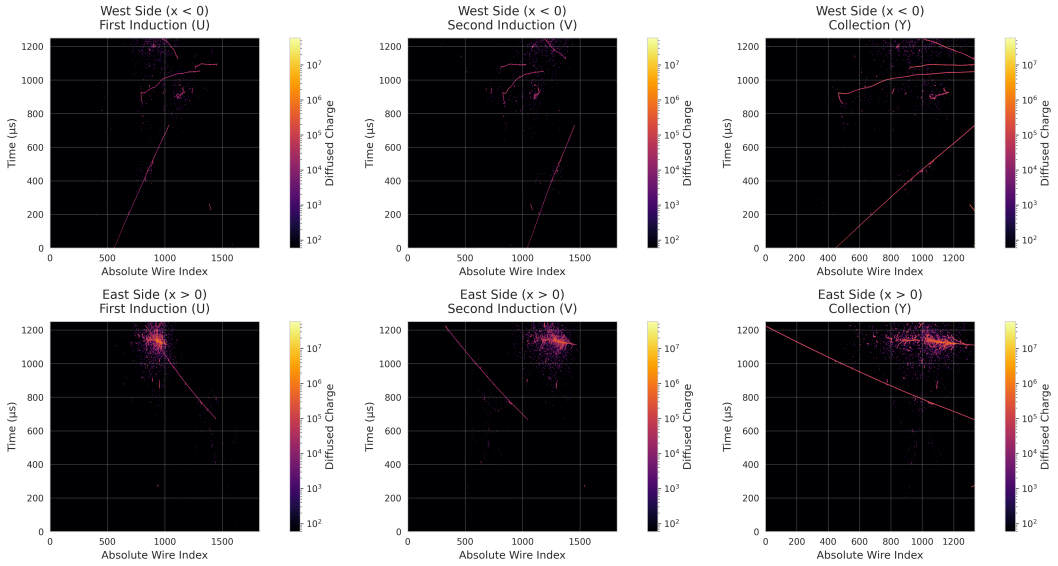


Figure 2: **Diffused Charge Projections.** Six 2D detector views of the same 3D event after electron drift and diffusion. Top row: three wire plane projections from the west detector side. Bottom row: three projections from the east side. Each plane captures the event at a different angle (U: +60°, V:-60°, Y: 0°), with axes showing wire channel number vs. drift time in microseconds.

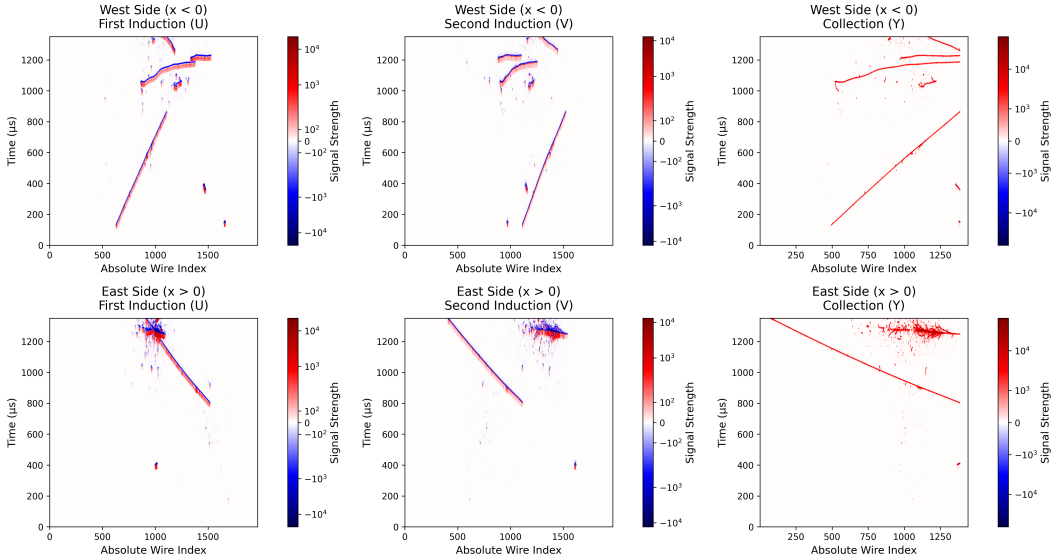


Figure 3: **Final Detector Readout.** The same six views after full detector simulation including field response and electronics effects. This represents the actual sensor data that reconstruction algorithms must process. Induction planes (U, V) show induced signals while collection planes (Y) show collected charge.

receives as input, requiring models to work backwards through multiple layers of physics to recover the original 3D particle information.

The reconstruction challenge is hierarchical. At the signal level, algorithms must process noisy, shaped waveforms. At the geometric level, they must solve the tomographic reconstruction from limited angular projections, complicated by the fact that the two detector sides see different 3D volumes due to the finite drift length. At the semantic level, models must identify particle types and interaction topologies from the reconstructed patterns.

Data Component	Dimensions	Storage
3D Segments	On the order of 10^5-10^6 segments per event 0.1 mm spatial resolution	$\sim$5 MB
2D Wire Projections Induction (U) Induction (V) Collection (Y)	6 total planes 2 planes × 1970 channels × 2700 time samples 2 planes × 1970 channels × 2700 time samples 2 planes × 1440 channels × 2700 time samples 3 mm wire spacing, 0.5 μs time sampling Extreme sparsity with <2% of the images active	$\sim$10 MB 5 MB waveforms 5 MB 2D truth
Optical Signals	162 photosensors × 12,960,000 time samples 1 ns time resolution, can be reduced greatly by sparsification	<5 MB
Total per Event		**$\sim$20 MB**

Table 1: Dataset dimensions per event. The combination of high-resolution 3D segments and multiple 2D projections provides rich supervision for learning detector physics.

B Targeted Data Challenges

There are three classes of data challenges and AI research opportunities associated with this dataset, including tomographic reconstruction, 3D pattern recognition, and foundation models (FMs).

B.1 Tomographic Reconstruction

The task is to infer the 3D scene of particle trajectories (point cloud) from three projected 2D images with different projection angles. The quality is measured by comparing against the labels, the 3D point clouds of energy depositions (i.e. input to the detector simulation). The geometric distance between the inferred and label points is a primary metric for evaluating quality. Secondly, for all pairs of matched between inferred and label points, a difference in the estimated energy deposition is also used as an equally crucial metric as the geometric distance. Third, inference speed and scalability is an important metric to ensure the developed techniques can be used for a larger detector such as the DUNE far detector.

For this challenge, we plan to provide a subset of 10M images with labels. The size of the subset is to be determined. For a reference, the machine (deep) learning models used in the community are typically trained using $\mathcal{O}(100k)$ images, and we plan to publicize the labels for no more than 200k subset until the data challenge is over.

B.2 3D Pattern Recognition

There are multiple inference tasks. Machine lerning models for these tasks may be optimized via supervised learning. The labels will be provided for the same subset of data described for the tomographic reconstruction.

Keypoint detection: A particle travels along the recorded trajectory of the particle. Knowing the start and end points is crucial to infer the direction of travel of the particles and their correlations with other particles. The metrics are computed for every matched pair of label and inferred points. The matching is produced based on the geometric distance. The metrics include the mean distance between all of the label and inferred points, 50%, and 90% quantile.

Panoptic Segmentation: Pixels must be partitioned to infer the semantic types at three different levels of fidelity. The lowest level fidelity is a semantic segmentation for different particle types. The second level distinguishes individual particle instances. The highest level is concerns the *interactions* which represents a group of particle instances that share the same creation physics origin. The quality is measured for those three levels of semantic fidelity separately using the Adjusted Rand Index (ARI) as a common metric for clustering.

Particle Flow: Within each inferred interaction, directed relation of particle instances must be inferred. There may be multiple *primary* particles that are created simultaneously at the interaction

origin. Some particles may be decay products of other particles. Such parent-child relationship can be described in a directed graph, and the metrics will compare the similarity of the label and inferred graphs.

B.3 Foundation Models

The scale (10M images) of the dataset will critically enable the development of the FMs. Techniques for effective representation learning is crucial knowledge yet to be mastered for the subject detector technology. The challenge task is to develop an effective representation learning for pre-training. The model will be then tested against the tomographic and pattern recognition challenge tasks to demonstrate how well the pre-training worked to extract general features. This challenge can utilize all dataset, and fine-tuning may be done using a subset of labeled datasets.