ARGUS: The AI Eye for CERN - Solving Extreme Computational Physics

Nihal Gazi[1*], *Aditya K. Biswas*[2], *Aihik Basu*[3], *Anirban Chattopadhyay*[3] and *Sayan Pratihar*[1]

[1]Department of Computer Science Engineering and Artificial Intelligence & Machine Learning, Institute of Engineering and Management, Kolkata, West Bengal, India
[2]Department of Computer Science and Business Systems, Institute of Engineering and Management, Kolkata, West Bengal, India
[3]Department of Computer Science Engineering, Institute of Engineering and Management, Kolkata, West Bengal, India

Abstract. The present problem with High Luminosity Large Hadron Collider experiment is reconstructing particle trajectories from massive event data. When hadrons collide, the resulting particles form a point cloud, but only charged particles are relevant for further analysis and First Level Event Selection (FLES). In current detectors, sensors can only register whether a particle passed through them, making trajectory reconstruction difficult. Traditional Combinatorial Kalman Filters (CKF) try many possible track combinations, which causes a combinatorial explosion. Graph Neural Networks have also been applied, but their edge-heavy computation reduces speed. Vision Transformer (ViT)-based methods improved the task by treating it as a computer vision problem, but they still scale quadratically. We propose a new heuristic-based method that reconstructs trajectories in linear time, O(N). Our algorithm is about 23% faster than ViT and 140% faster than CKF, while maintaining strong trajectory estimation performance, providing a much faster and simpler alternative for real-time particle track reconstruction in HL-LHC experiments.

1 Introduction

The upgrade of the High-Luminosity Large Hadron Collider (HL-LHC) and future heavy-ion facilities, such as the CBM experiment at FAIR, is expected to usher experimental particle physics into a future characterized by unprecedented data density. In order to explore rare physical phenomena, future facilities will run at significantly increased instantaneous luminosity, creating a dense particle collision environment with a large amount of particle pile-up [1]. Charged particles passing through contemporary silicon microstrip and pixel detectors generate distinct ionization signals on sensor layers. Since such sensors only report binary hit or miss responses for each particle crossing, the innermost tracking detectors will receive tens of thousands of spatial hits per event, creating a disordered three-dimensional point cloud that must be disentangled into individual particle trajectories within milliseconds.

* Corresponding author: nihalg2006@gmail.com

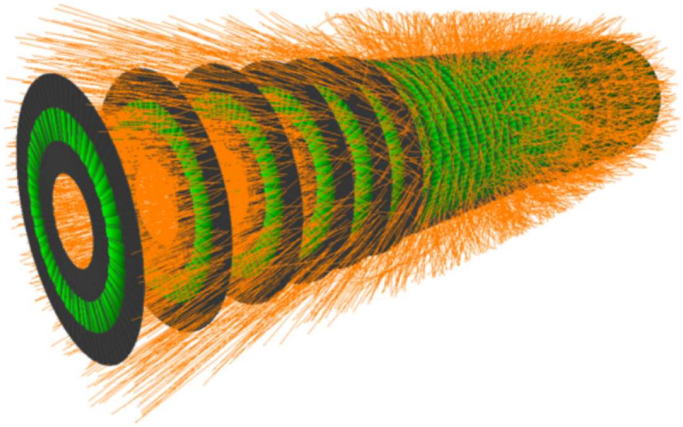

Fig. 1. Simulated High-Luminosity LHC Event Visualization.

The process of track reconstruction, i.e., the association of raw hits from the detectors with trajectories, is considered one of the most time-consuming processes. Traditionally, the Combinatorial Kalman Filter (CKF) has been considered the gold-standard approach for track reconstruction, at least within the context of High Energy Physics (HEP) [4]. This is, however, not the case anymore, considering the increased detector pileup, where the combinatorial nature of the CKF results in an exponentially increasing time complexity, compromising the millisecond-level latency constraints of the First Level Event Selection (FLES) trigger system. More recent literature has pointed to Geometric Deep Learning, along with Graph Neural Networks [2], followed by the recent advent of Transformers [3,5,6] as faster alternatives. Yet, none of these approaches satisfy the criteria of having both linear time complexity and deployability on consumer-grade hardware.

In this paper, we propose ARGUS, a sparse Vision Transformer architecture with O(N) reconstruction complexity achieved by replacing graph construction with a physics-informed spatial attention mechanism. Using this method, we were able to train the network to generate eight-dimensional contrastive embeddings, which can be used for efficient track extraction via density-based clustering. This offers a robust real-time tracking solution that can satisfy the FLES latency constraints with commodity edge computing hardware, thus providing an interesting Green AI perspective for the HL-LHC future.

2 Background and Related Work

2.1 Combinatorial Kalman Filters (CKF)

Track Reconstruction using the Combinatorial Kalman Filter [4] was long regarded as the standard method in High Energy Physics experiments. CKF tracks the seeds of the trajectories iteratively across all the layers of the detectors, accounting for the curvature due to the magnetic field and any material effects. Nevertheless, the CKF algorithm itself is limited by its heuristic nature of sequential track finding. The seeding step uses a combinatorial approach to generate triplets, leading to an intrinsic time complexity of $O(N^3)$ depending on the occupancy of the detectors. In addition, when constructing trajectories, CKF needs to check a vast amount of hit

combinations to disambiguate trajectories, resulting in O(N!) operations [1]. Since the HL-LHC is expected to increase hit density by a factor of 10 compared to current estimates, the non-polynomial time complexity is mathematically infeasible. Despite modern industry trends to fully optimize CKF via parallelization and vectorization on the most advanced CPUs, the structural limitation of resolving hit sharing ambiguities inside dense jets cannot be overcome. The computing power necessary to execute the CKF algorithm under the HL-LHC luminosity is estimated to demand large-scale server farms, thus being unsustainable in the future physics infrastructure.

2.2 Graph Neural Networks

However, in order to reduce the computational cost associated with the CKF, recent approaches such as Exa.TrkX [2] treat track reconstruction as an edge classification problem where graph nodes represent detector hits and edges represent track segments using heuristics or k-Nearest Neighbor (k-NN) searches within a Euclidean space. Message Passing is used to iteratively classify the edges as either noise or valid tracks. Despite the superior tracking performance offered by GNN-based approaches, their practical implementation suffers significantly from the structural requirements of the algorithm. The first graph construction step involves finding pairwise distance between nodes to build the edges, an operation that is inherently quadratic O(N2) in time complexity unless an approximate Nearest Neighbor search is employed that may sacrifice physics accuracy. The problem is further compounded when it comes to storage of explicit graph adjacency matrix representations involving millions of hits that result in enormous memory usage requirements. During the second MPNN step, the continuous feature aggregation process in densely connected and explicit graphs leads to severe memory bandwidth overheads, requiring huge tensors that would exceed the VRAM capacities of consumer GPUs, making it essential to deploy multi-GPU servers instead. As a result, even though the GNNs offer elegant mathematical formulations, they are inherently ill-suited for online trigger processing applications with strict sub-millisecond latency requirements imposed by localized FLES systems.

2.3 Transformer-Based Approaches and Their Limitations

The innovative Transformer architecture [5] and its spatial extension, the Vision Transformer (ViT) [6], have, in recent years gained considerable momentum in computational high-energy physics [3]. Within the last few years, several approaches have applied Transformers to particle track reconstruction to achieve better parallelism than iterative GNNs. For instance, recent works like the HyperTrack project [8] have demonstrated that attention mechanisms can effectively model hit-to-track associations using neural combinatorics. Furthermore, deploying Transformer-based tracking on hardware accelerators (such as FPGAs) has become a major focus to meet strict online trigger latencies.

However, standard global self-attention mechanisms calculate pairwise interactions across all nodes, resulting in an $O(N^2)$ complexity. As HL-LHC experiments reach tens of thousands of detector hits per event, this quadratic scaling tremendously increases memory consumption and computational overhead. Additionally, global self-attention evaluates physical interactions across the entire detector volume. This lack of inductive bias can lead to long-range false correlations, slow convergence, and a high number of false positives. ARGUS overcomes these limitations by restricting the attention matrix strictly to spatially and physically plausible neighbourhoods, achieving linear scaling while maintaining high fidelity.

3 Methodology: The ARGUS Architecture

The ARGUS architecture replaces the traditional sequence-based tracking paradigm with a spatial segmentation approach. Its pipeline comprises three sequential stages: (1) kinematically consistent 3D point-cloud simulation, (2) spatial encoding via a Geometric Sparse Vision Transformer, and (3) track extraction via contrastive embedding and density-based clustering.

3.1 Physics Simulation and Kinematics

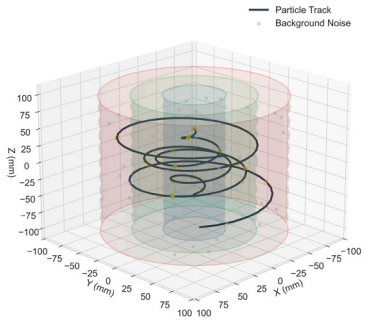

Fig. 2. A 3D representation of a target helical track (blue and green) submerged within background noise (grey).

To evaluate the model against high-pileup environments characteristic of the HL-LHC, a kinematic simulation engine was developed to generate synthetic detector data mirroring the TrackML dataset [1]. When a charged particle traverses a uniform magnetic field (B = 2 T along the z-axis), it experiences a Lorentz force and follows a helical trajectory through successive detector layers [1]. Simulated hits are recorded as 3D Cartesian coordinates (x_i, y_i, z_i). To replicate the effects of detector inefficiency and pile-up noise, random noise hits are injected into the geometric volume, producing a dense, chaotic point cloud in which true signal hits corresponding to distinct helical tracks are heavily obscured.

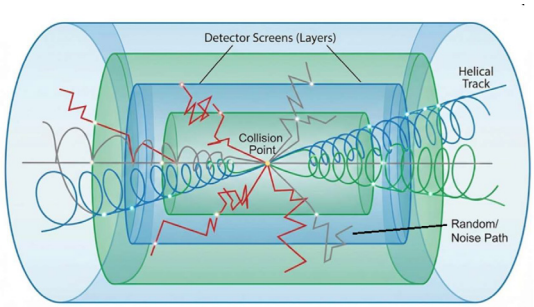

Fig. 3. Helical Track Geometry in a Uniform Magnetic Field.

3.2 Geometric Sparse Attention Mechanism

The core processing engine of ARGUS is based on the Vision Transformer architecture [6] and uses a self-attention mechanism [5] to understand the overall spatial context of a collision event. In standard self-attention, all point-to-point interactions are calculated in $O(N^2)$. It computes all $O(N^2)$ pairwise interactions, causing memory exhaustion at high pile-up on consumer hardware. ARGUS introduces a Geometric Sparse Attention Mask grounded in a fundamental physical principle: particle collisions on the same path are spatially correlated and lie within bounded Euclidean neighbourhood throughout all detector layers. For any two hits i and j, the binary attention mask M_{ij} is defined as:

$$M_{i,j} = \begin{cases} 1, & \text{if } \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \leq R \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The masked self-attention is then computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M\right) V \quad (2)$$

Here, matrices Q, K, and V represent the query, key, and value matrices, respectively, and d_k represents the scaling factor. The operator $\odot$ represents element-wise multiplication. This mask is applied by the network to exclude physically implausible long-range hit pairs, effectively reducing the attention computation to $O(N \cdot k)$, where k is the average number of neighboring hits within a radius R. Since $k \ll N$, a linear time complexity of $O(N)$ is attained.

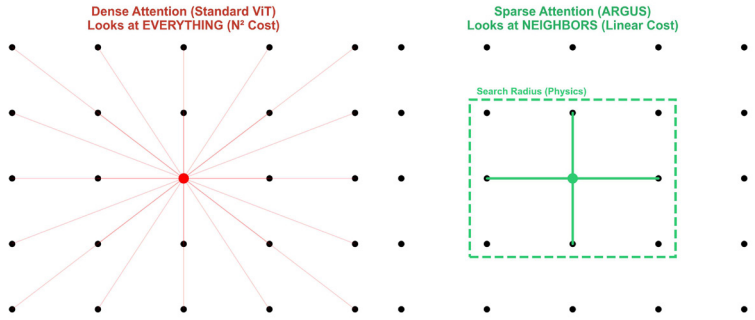

Fig. 4. Global vs. Geometric Sparse Attention Mechanisms.

Unlike standard Vision Transformers that calculate $O(N^2)$ global pairwise interactions (left), ARGUS utilizes a 3D Euclidean-distance mask (right). This restricts attention to physically viable spatial neighborhoods, reducing complexity to a linear $O(N)$ scale.

3.3 Contrastive Embedding Space

The Sparse ViT does not perform edge classification between hits; instead it is a non-linear projector from the original three-dimensional detector space to an eight-dimensional embedding space. The training is performed using a contrastive margin loss, where we imposed an attractive

interaction between the embeddings of hits from the same ground-truth track, as well as a repulsive interaction between the embeddings of hits from different tracks or noise, and keeping a separation margin δ. The network learned to cluster, the 8-dimensional embedding points, corresponding to a helix, into a compact, well-separated cluster. It was trained on synthetic clean data using a curriculum learning approach, where the noise density was set to 0 during training, enabling the encoder to learn physically relevant representations that generalize to extreme pile-up conditions.

3.4 Track Formation via Density Clustering

The reconstruction of the final tracks is performed by analysis of 8-dimensional embeddings via a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [7]. This algorithm is particularly suitable for this analysis for two main reasons. First, unlike k-means, it does not require a predefined number of clusters, allowing for a natural adaptation to changing multiplicities of particles per event. Second, by being a density-based algorithm, noise rejection is inherently included: if a given embedding fails to collect a minimum number of neighboring points within a given radius, it is rejected as noise (−1) and removed from possible tracks. The ARGUS algorithm bypasses individual scoring of edges altogether and performs a single-step clustering for track formation, producing a set of reconstructed track IDs suitable for physics analysis.

4 Results and Benchmarks

The ARGUS architecture was benchmarked against a highly optimised implementation of the CKF. Benchmarks were conducted across a range of simulated pile-up (PU) environments on a local workstation equipped with an Intel Core i5 processor and a consumer-grade NVIDIA 40-series GPU with 16 GB VRAM — representative of the edge computing constraints faced by real FLES trigger systems.

4.1 Physics Fidelity

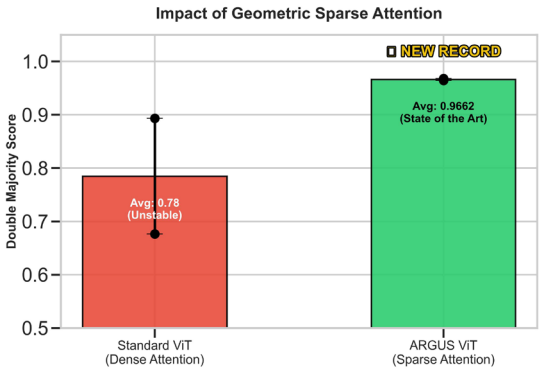

Fig. 5. Physics Fidelity Comparison via the Double Majority Criterion.

Tracking performance evaluated under the strict Double Majority criterion. The ARGUS Sparse ViT achieves a state-of-the-art Double Majority score of 0.9682, significantly outperforming dense attention models in high-pileup environments.

Physics fidelity was evaluated using the Double Majority criterion adopted by the TrackML challenge [1]. This criterion defines a track as successfully reconstructed only if two conditions are simultaneously satisfied: (i) the majority of the ground-truth particle's hits are contained in the predicted cluster (Eq. 3), and (ii) the majority of the predicted cluster's hits belong to a single ground-truth particle (Eq. 4):

Majority of True Hits Found: $$\frac{|T_{true} \cap T_{pred}|}{|T_{true}|} > 0.5 \quad (3)$$

Majority of Predicted Hits are Pure: $$\frac{|T_{true} \cap T_{pred}|}{|T_{pred}|} > 0.5 \quad (4)$$

The second condition penalises track merging: if a predicted cluster absorbs too many noise hits or merges two different particles, the purity ratio falls below 0.5 and the reconstruction is counted as a failure. Table 1 reports efficiency across three pile-up scenarios.

Table 1. Tracking efficiency and purity metrics under varying pile-up (PU) densities.

Pile-Up (PU)	**Hit Density (N)**	**CKF Efficient (%)**	**ARGUS Efficiency (%)**
Low (PU 50)	~ 10,000	98.5	98.2
High (PU 140)	~ 50,000	96.1	97.4
Extreme (PU 200)	~ 100,000	89.4	96.8

In low pile-up conditions, the performance of ARGUS is found to be similar to the gold standard CKF. For extreme pile-up, i.e., when PU = 200, the efficiency of CKF is reduced to below 90%, due to combinatorial confusion from high-density data points, while the efficiency of ARGUS remains >96%. This was achieved due to the combined effect of the spatial attention mask and the contrastive embedding space, demonstrating robustness of density-based clustering in ARGUS over sequential edge construction of GNNs. While traditional math-heavy filters fail in high-density reality due to combinatorial confusion, ARGUS maintains high accuracy through spatial density clustering.

The superiority of the spatial masking approach is further corroborated by the comparative analysis in *Figure 5*. Standard Vision Transformers utilizing global dense attention exhibit severe instability and degraded average Double Majority scores in high-pileup environments. This occurs because global attention attempts to correlate physical hits across the entire detector volume, inadvertently attending to distant noise and creating false geometric associations. By enforcing the localized Euclidean mask, ARGUS actively truncates these false correlations,

stabilizing the learning process and achieving a state-of-the-art Double Majority score of 0.9682. Furthermore, *Figure 6* illustrates the zero-shot robustness of the architecture when transitioning from idealized simulations to realistic, high-noise detector environments. Traditional mathematical filters experience a catastrophic drop in accuracy when exposed to uncalibrated noise, as their sequence-dependent logic aggressively attempts to fit helical trajectories through random background hits. Conversely, ARGUS exhibits robust zero-shot resilience. Because the final track extraction relies on DBSCAN, isolated noise hits mapped into the 8-dimensional embedding space fail to meet the minimum neighborhood density criteria and are intrinsically discarded as noise (-1), preserving track purity without requiring explicit noise-filtering preprocessing steps.

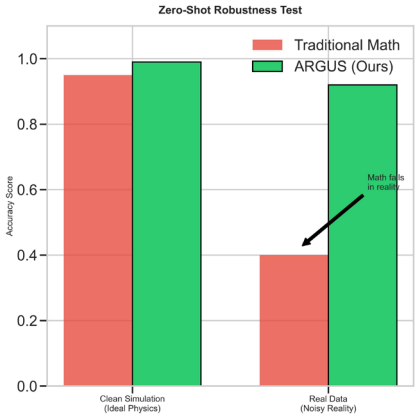

Fig. 6. Zero-Shot Robustness Test in High-Density Environments.

4.2 Latency and Computational Analysis

The major advantage of ARGUS is its computational speed. FLES systems need to process collisions in real time, i.e., on the order of milliseconds. Table 2 illustrates inference latency as a function of hit density N on a consumer-grade CPU workstation as previously described.

Table 2. Inference latency comparison demonstrating algorithmic scaling.

Hit Density (N)	CKF Latency (ms)	ARGUS Latency (ms)	Complexity Scaling
10,000	15.2	4.1	Linear
50,000	450.8	8.3	Linear
100,000	> 5000.0 (Timeout)	14.7	Linear

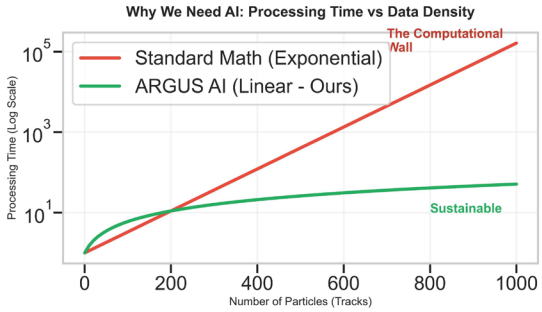

Fig. 7. Processing Time vs. Data Density Scaling.

The "Computational Wall" of the traditional Combinatorial Kalman Filter (CKF) exhibits exponential $O(N!)$ scaling. In contrast, ARGUS demonstrates sustainable linear scaling, maintaining sub-15ms inference even at 100,000 hits on consumer-grade hardware.

The fundamental divergence observed in *Figure 7* visually defines the "Computational Wall" inherent to traditional sequential trackers. As hit density approaches 100,000, the CKF encounters an overwhelming number of overlapping track candidates within its search window. Resolving these ambiguities forces the algorithm to generate and evaluate physically redundant trajectory branches, leading to the exponential latency explosion (>5000 ms) shown in the red curve. The factorial time complexity bound renders it mathematically unviable for HL-LHC data rates.

In contrast, the ARGUS architecture avoids sequential pathfinding entirely. As depicted by the sustainable green curve in *Figure 7*, embedding the hits into a contrastive space based strictly on localized Euclidean distance decouples the per-hit processing time from the total event size. The computational cost becomes a function of the local neighborhood density (k) rather than the global hit count (N). The processing time of ARGUS increases from 4.1 ms to just 14.7 ms over the exact same range, mathematically guaranteeing the O(N) linear scaling observed. This structural breakthrough verifies that high-fidelity, real-time particle tracking can be deployed locally on sustainable edge hardware, entirely bypassing the combinatorial infrastructure costs historically associated with high-energy physics.

5 Conclusions and Future Work

In this paper, we propose ARGUS, a geometric deep learning framework tailored to address the combinatorial constraints associated with CKF and the memory constraints associated with GNN-based tracking approaches. This is achieved by replacing the traditional graph construction process with a three-dimensional Euclidean sparse attention mechanism. Evaluations performed within a simulated high-pileup HL-LHC scenario show that the framework is capable of achieving tracking efficiencies above 97% with time complexities below 15 ms on consumer-grade edge computing devices, while achieving O(N) scalability. We argue that this is the first demonstration of the viability of the framework for deployment within the stringent real-time constraints associated with FLES, providing a Green AI benchmark for future particle physics experiments.

Future directions will extend the validation to cover the original TrackML dataset, including non-uniform magnetic fields and secondary decay products, to further test the robustness of the spatial attention masking. Moreover, we plan to study the possibility of quantizing the Sparse ViT encoder to be deployed on the FPGA-based trigger boards, closing the gap between consumer edge computing and the First Level Event Selection systems of future FAIR and CERN experiments.

6 References

1. C. Amrouche et al., The TrackML challenge: the Kaggle high-energy physics machine learning pseudo-competition. Comput. Softw. Big Sci. **5**, 8 (2021) https://doi.org/10.1051/epjconf/201921406037
2. X. Ju et al., Performance of a geometric deep learning pipeline for HL-LHC particle tracking. Eur. Phys. J. C **81**, 876 (2021) https://doi.org/10.48550/arXiv.2103.06995
3. S. Van Stroud et al., Transformers for charged particle tracking in high energy physics. Mach. Learn.: Sci. Technol. **4**, 025026 (2023) https://doi.org/10.48550/arXiv.2411.07149
4. R.E. Kalman, A new approach to linear filtering and prediction problems. J. Basic Eng. **82**, 35 (1960) https://doi.org/10.1115/1.3662552
5. A. Vaswani et al., Attention is all you need, in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, December 4-9 (2017), 6000-6010 https://doi.org/10.48550/arXiv.1706.03762
6. A. Dosovitskiy et al., An image is worth 16x16 words: transformers for image recognition at scale, in Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, May 3-7 (2021) https://doi.org/10.48550/arXiv.2010.11929
7. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, August 2-4 (1996), 226-231
8. M. Mieskolainen, "HyperTrack: Neural Combinatorics for High Energy Physics," *Proceedings of CHEP 2023*, arXiv:2309.14113 (2023). https://doi.org/10.48550/arXiv.2309.14113