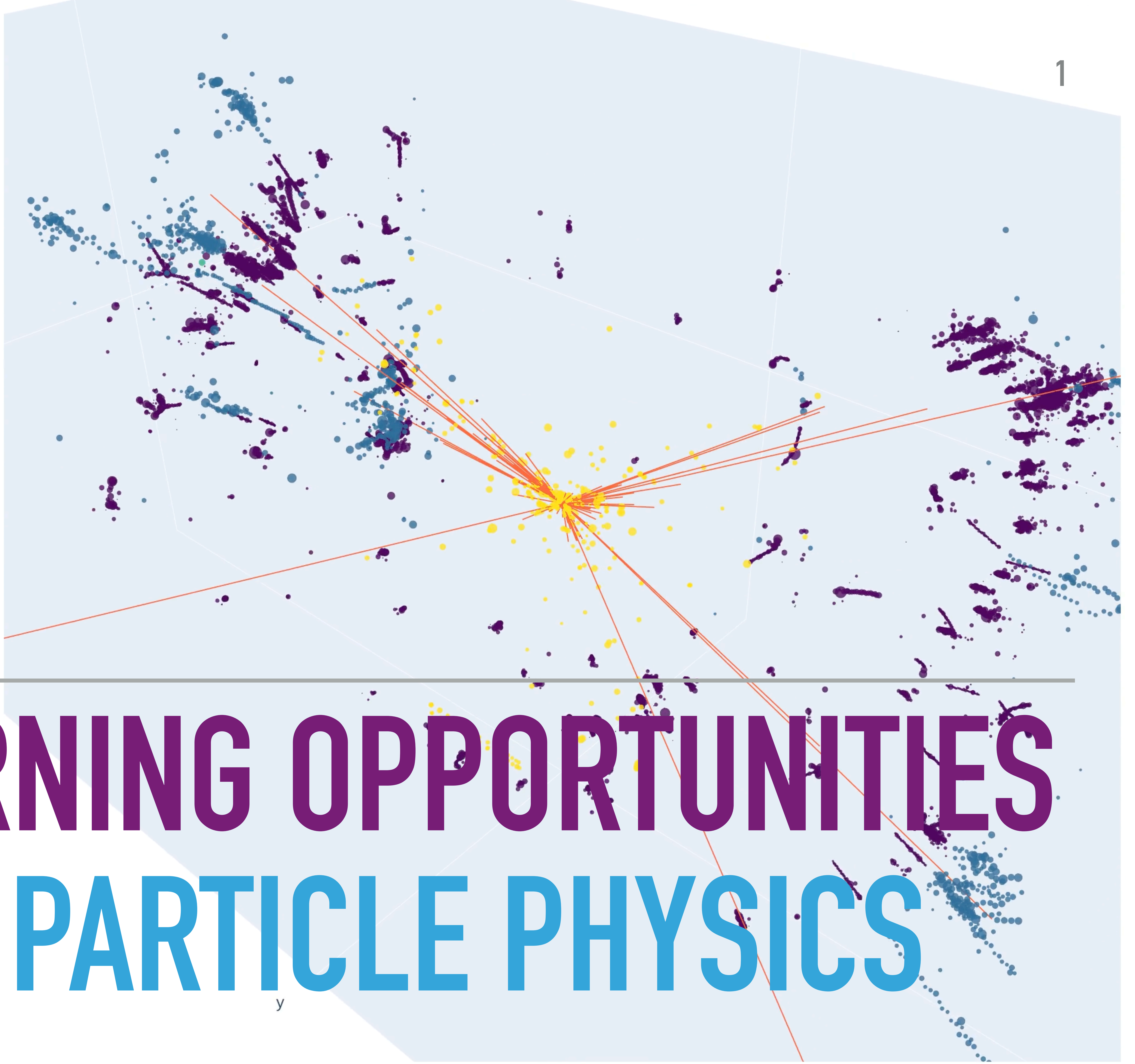


JAVIER DUARTE

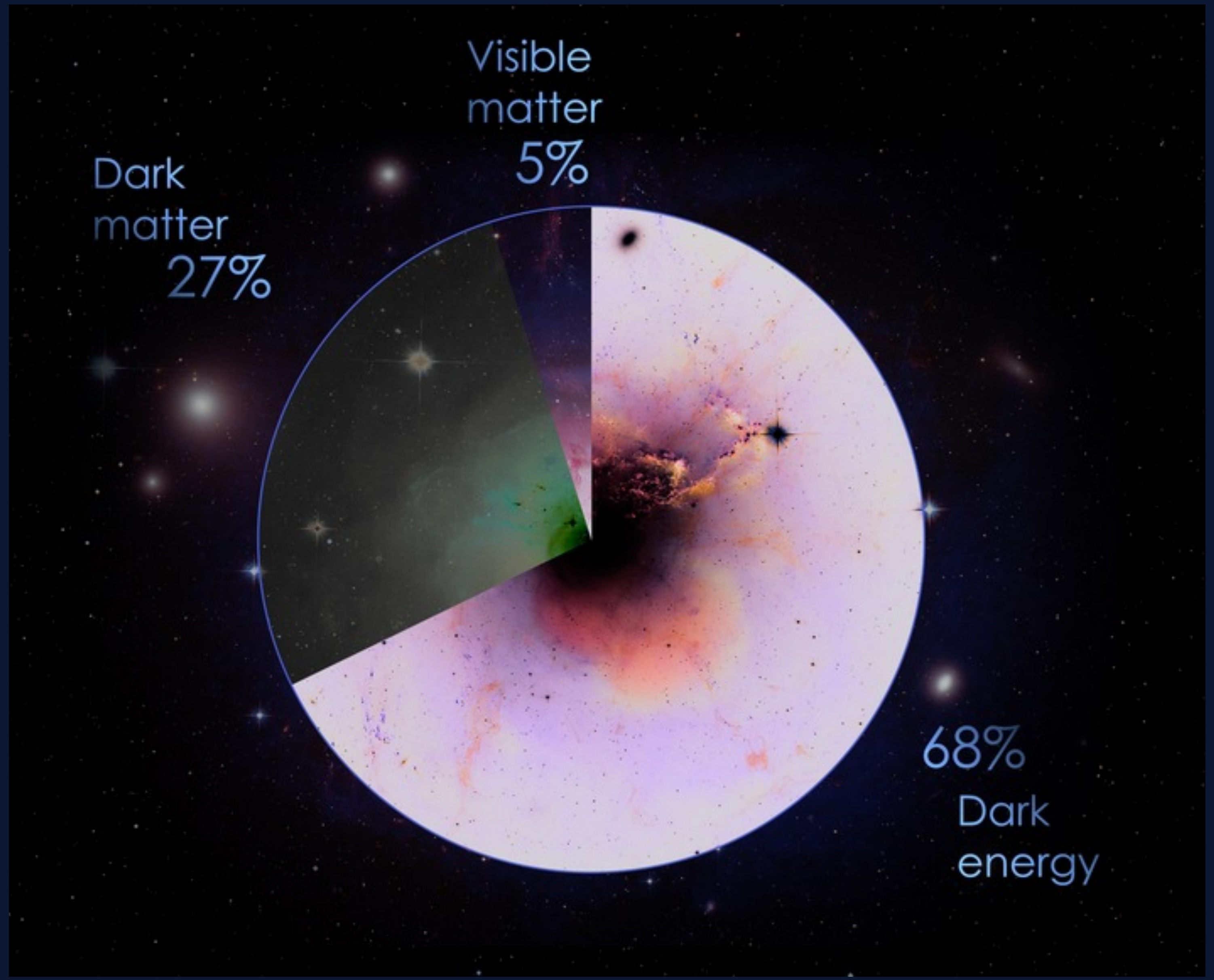
ICML

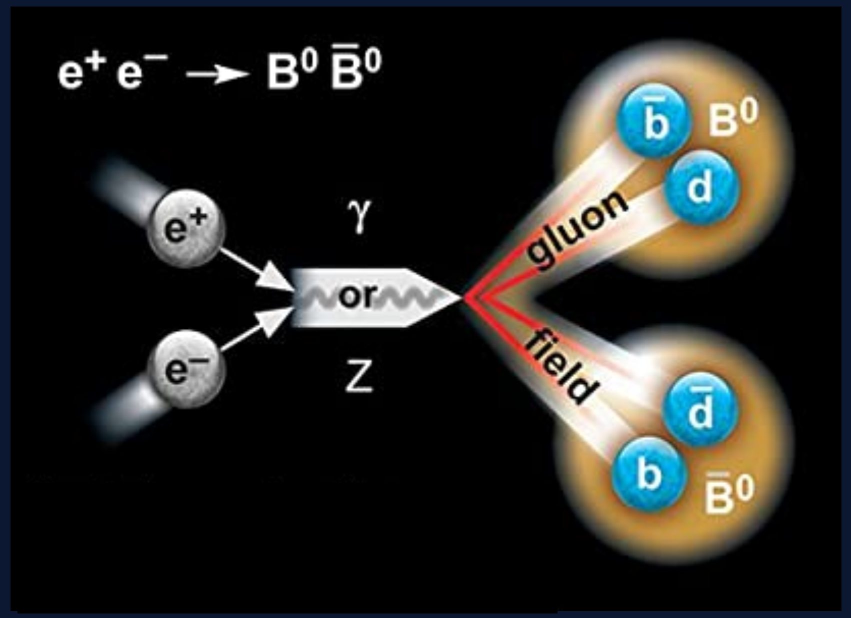
JULY 24, 2024

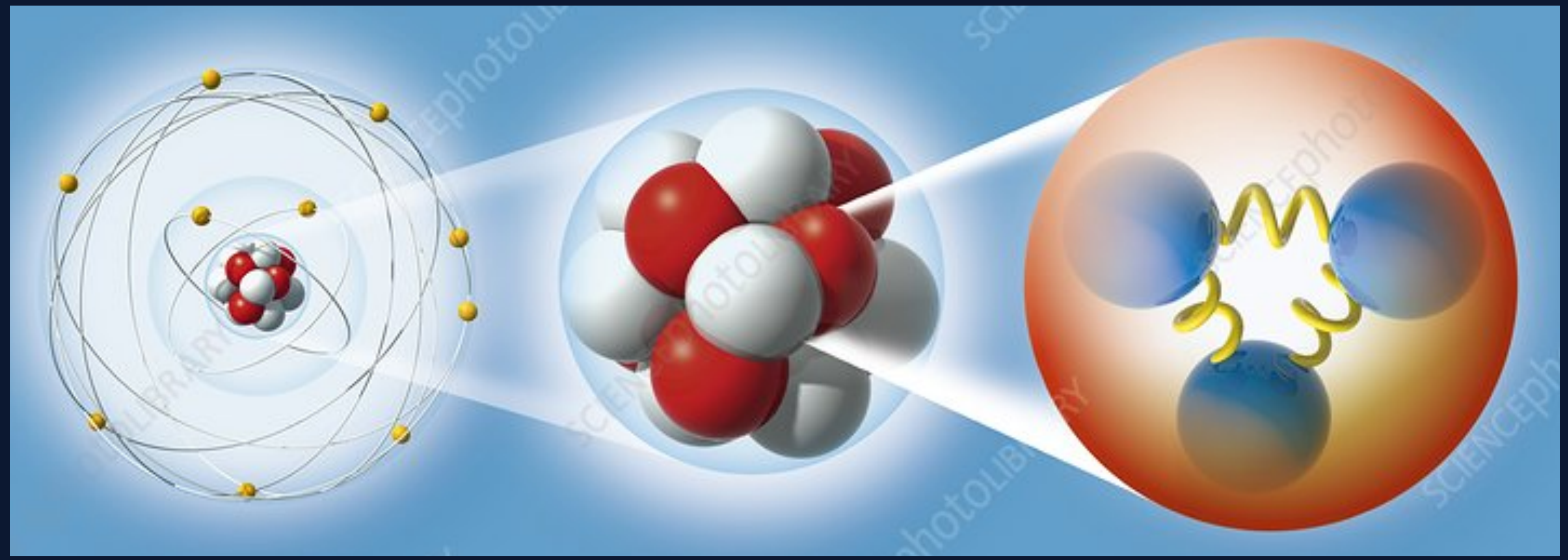
MACHINE LEARNING OPPORTUNITIES FOR NEXT GEN PARTICLE PHYSICS

BIG QUESTIONS

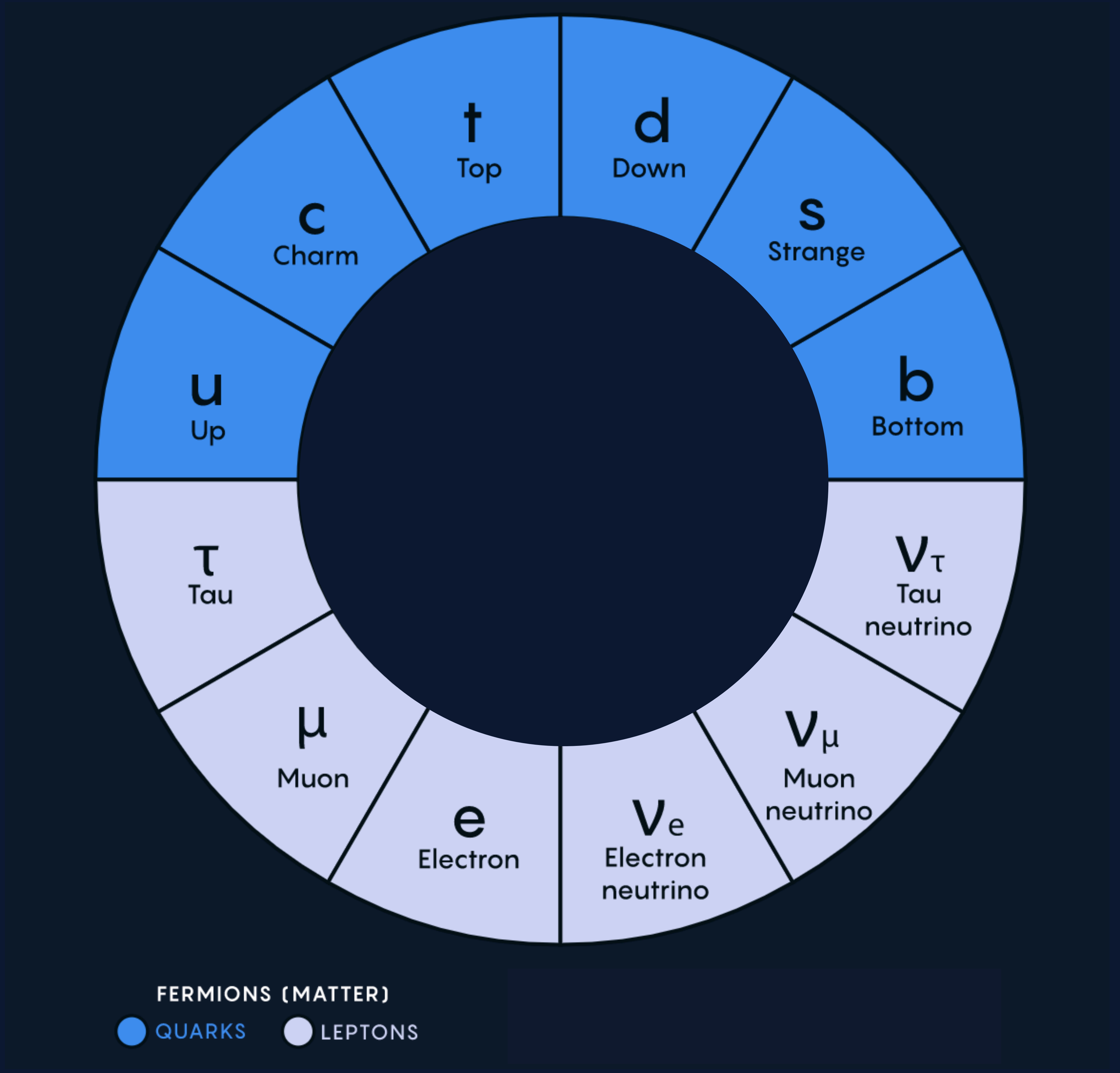
- What is our universe made of?
- What are the smallest building blocks of nature?
- How do they interact with each other?
- Is our universe stable?

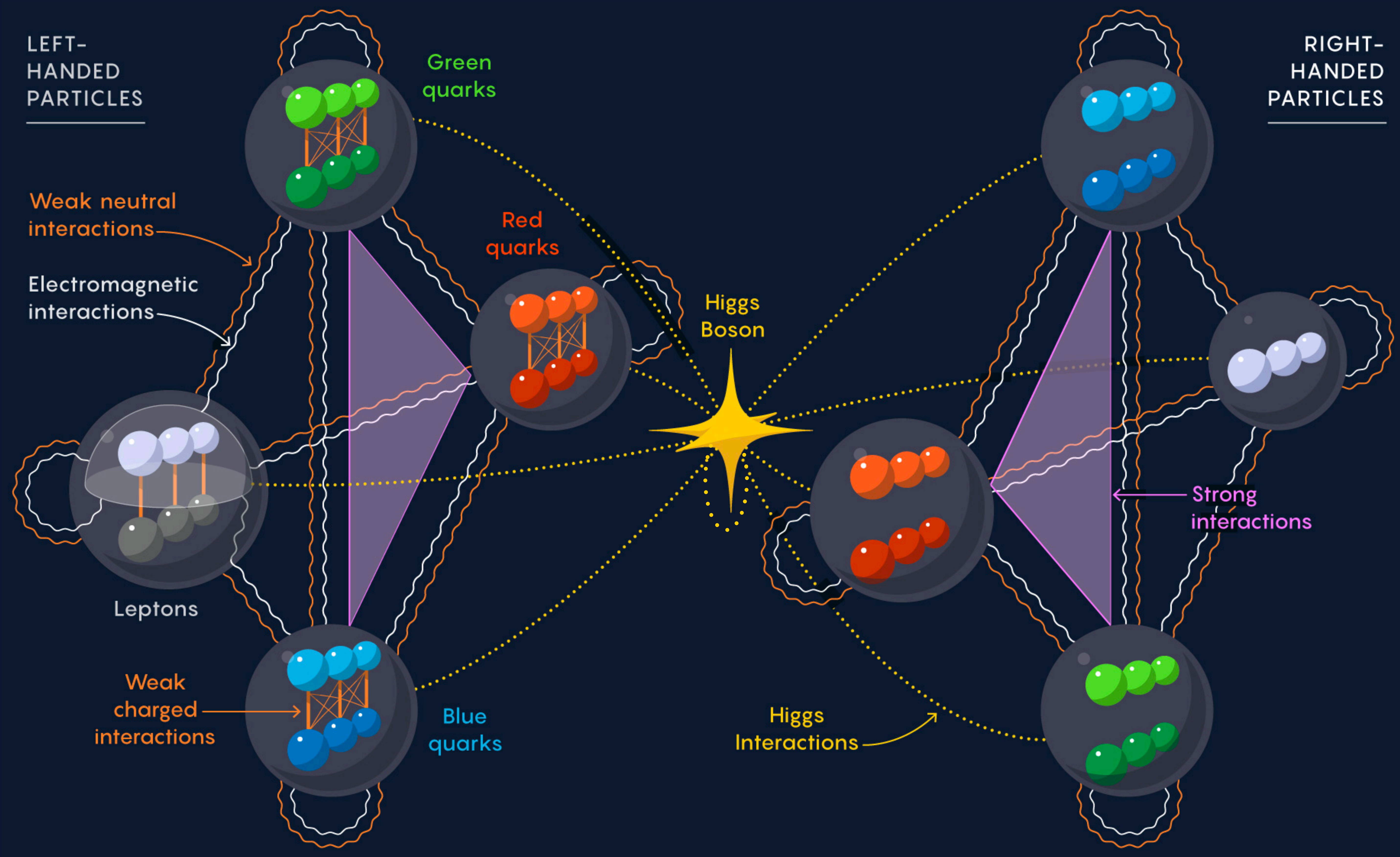






THE STANDARD MODEL





- But there has to be more to it! SM does not answer all our questions
- Higgs is a ***centerpiece:*** Mechanism by which particles ***acquire mass***
- How do we study these microscopic building blocks?

WHY HIGH ENERGY?

- High energies ⟷ short length & time scales
- Collisions at the highest energy possible today let us recreate conditions 0.1 ns after the Big Bang!

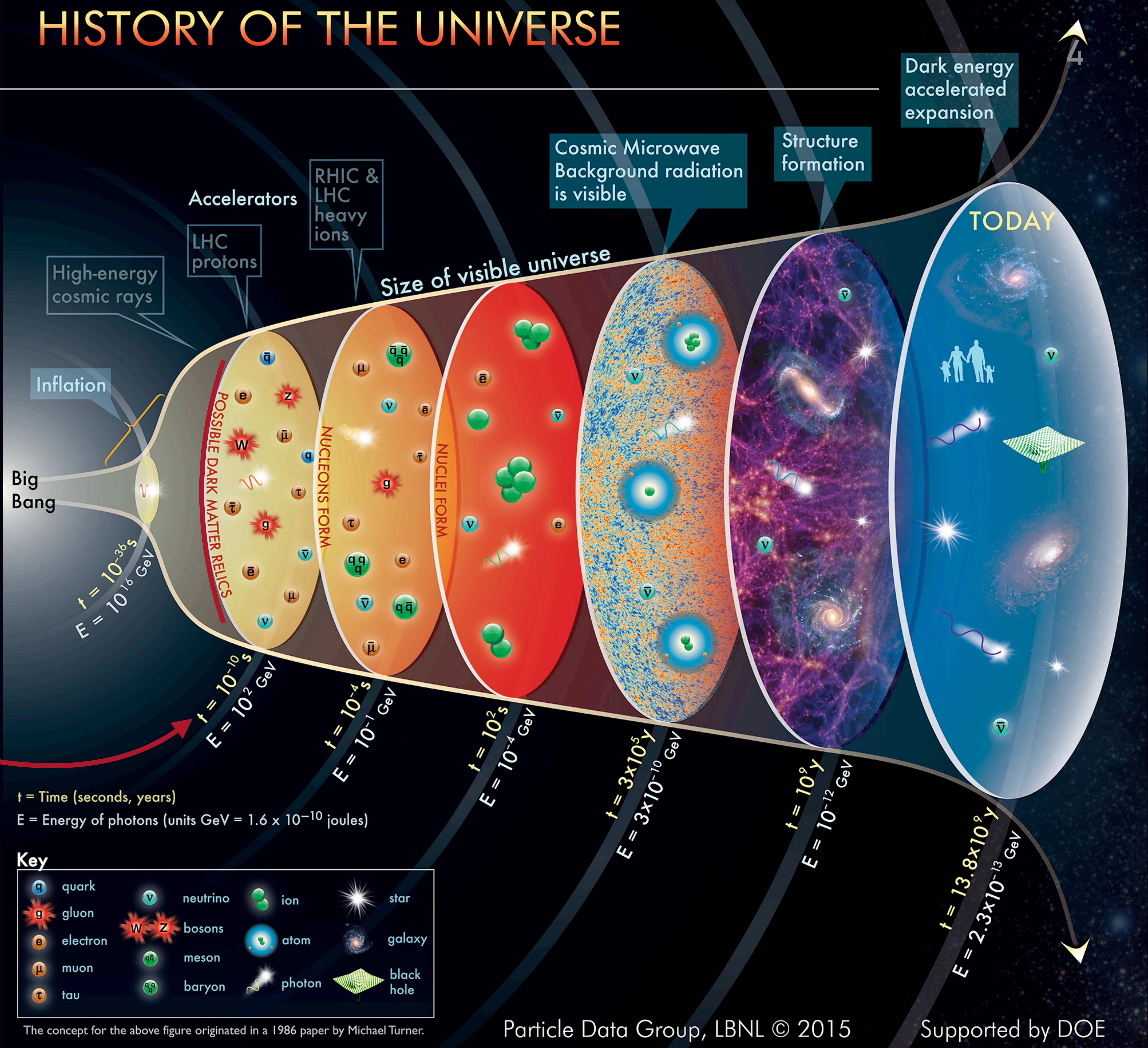
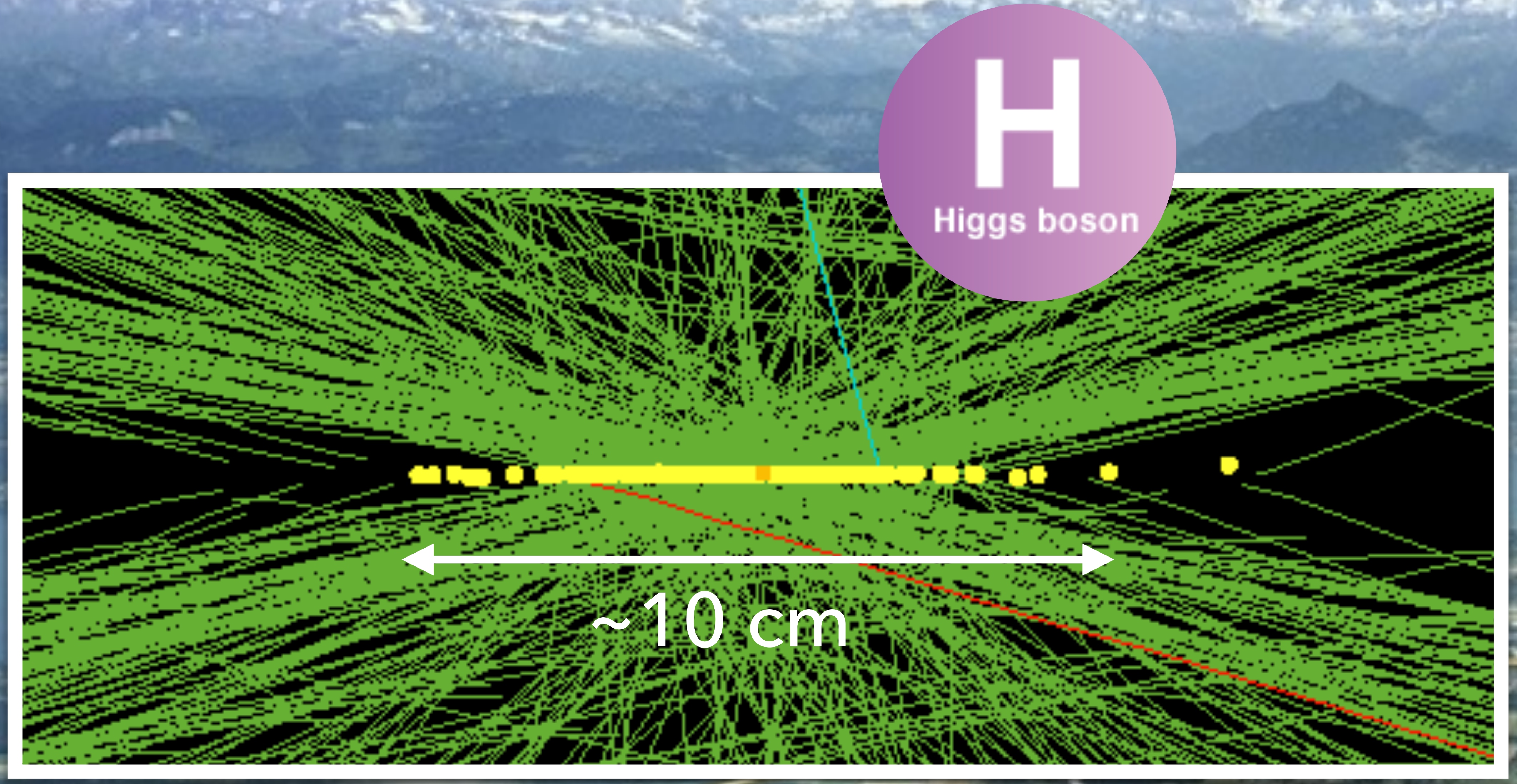
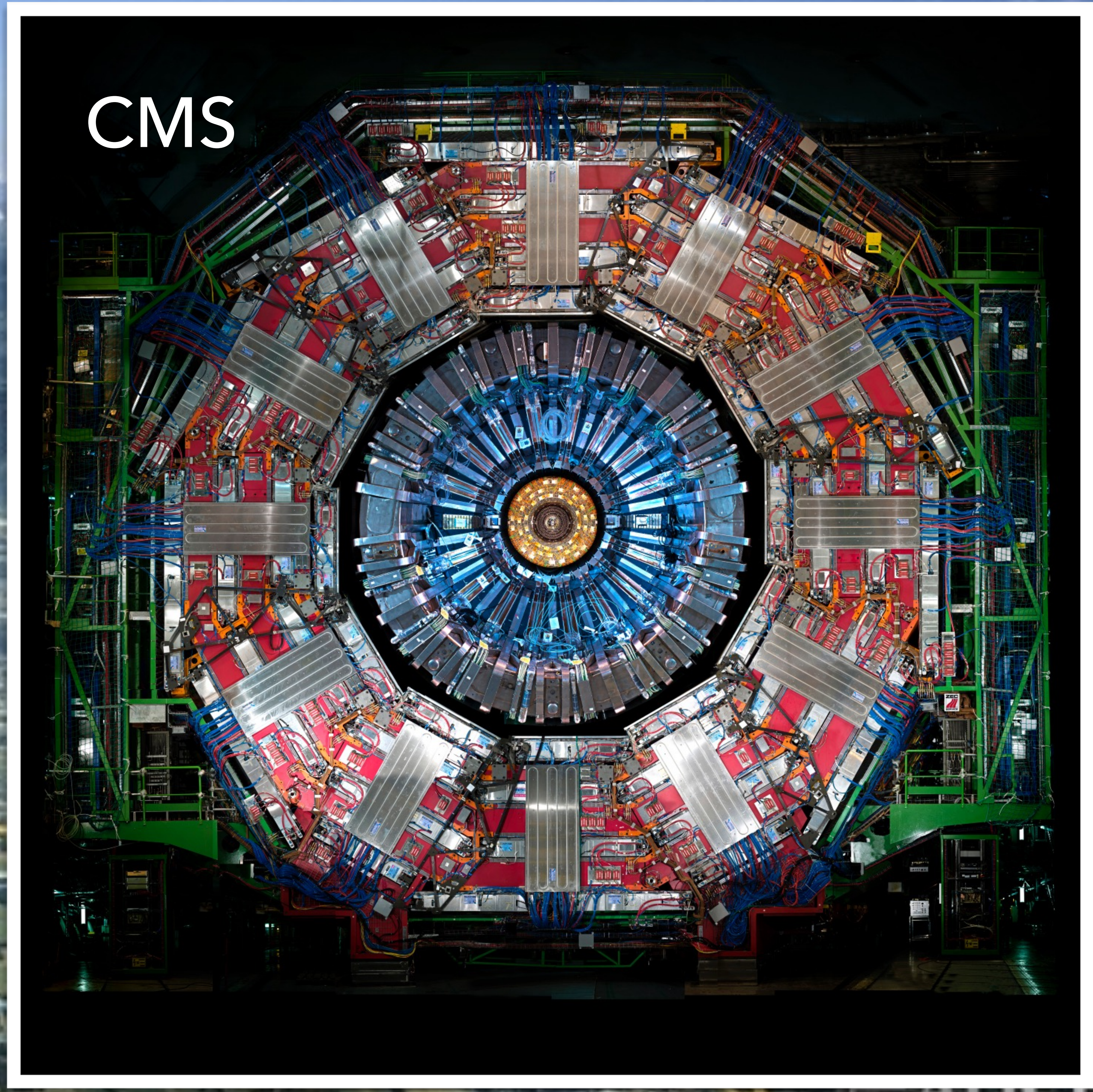


Image: https://particleadventure.org/history-universe

THE LARGE HADRON COLLIDER





proton-proton collider @ 13 TeV center-of-mass energy

4 interaction points

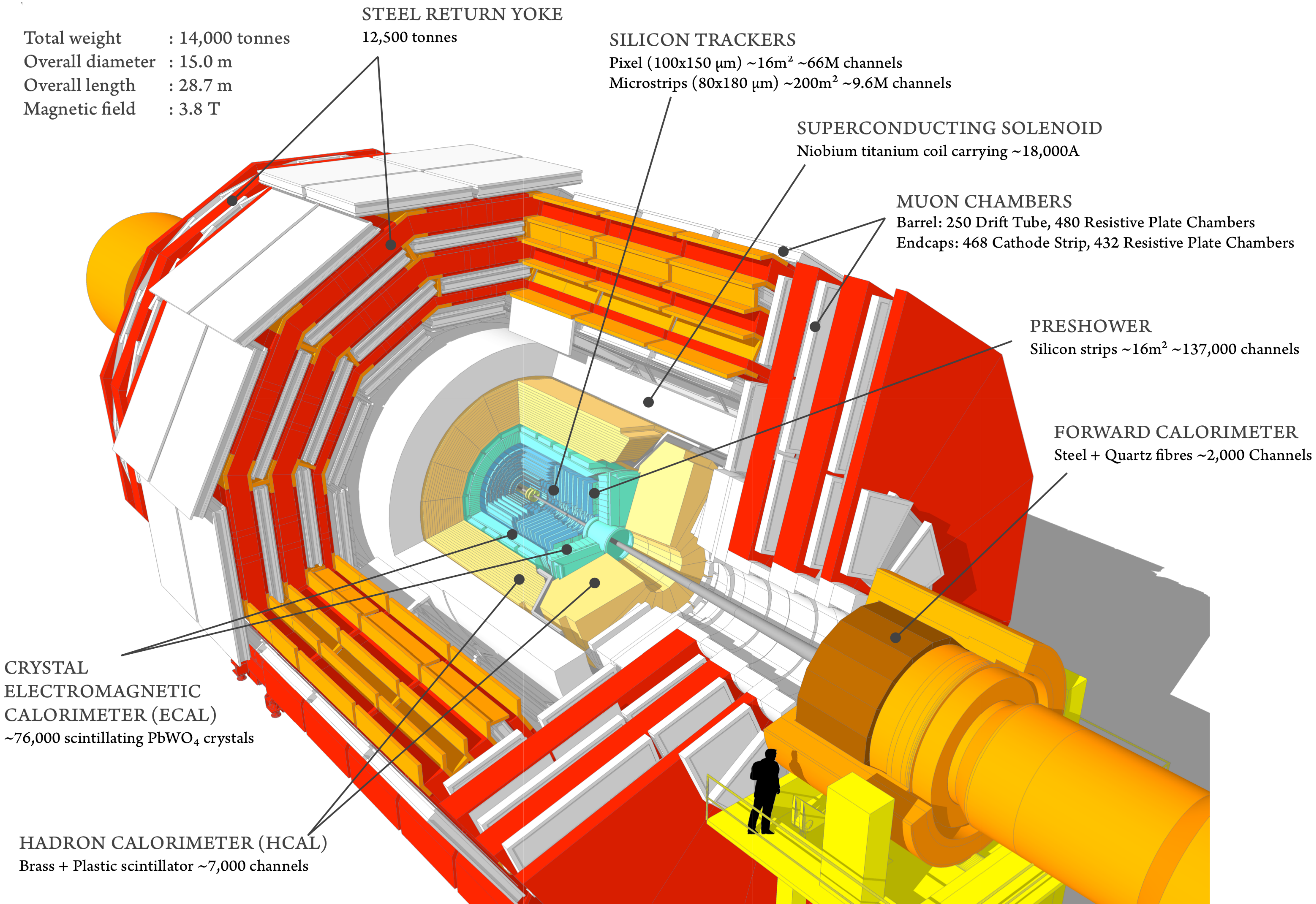
40 million collisions / second

Higgs boson produced 1/10 billion collisions (every 4 minutes)

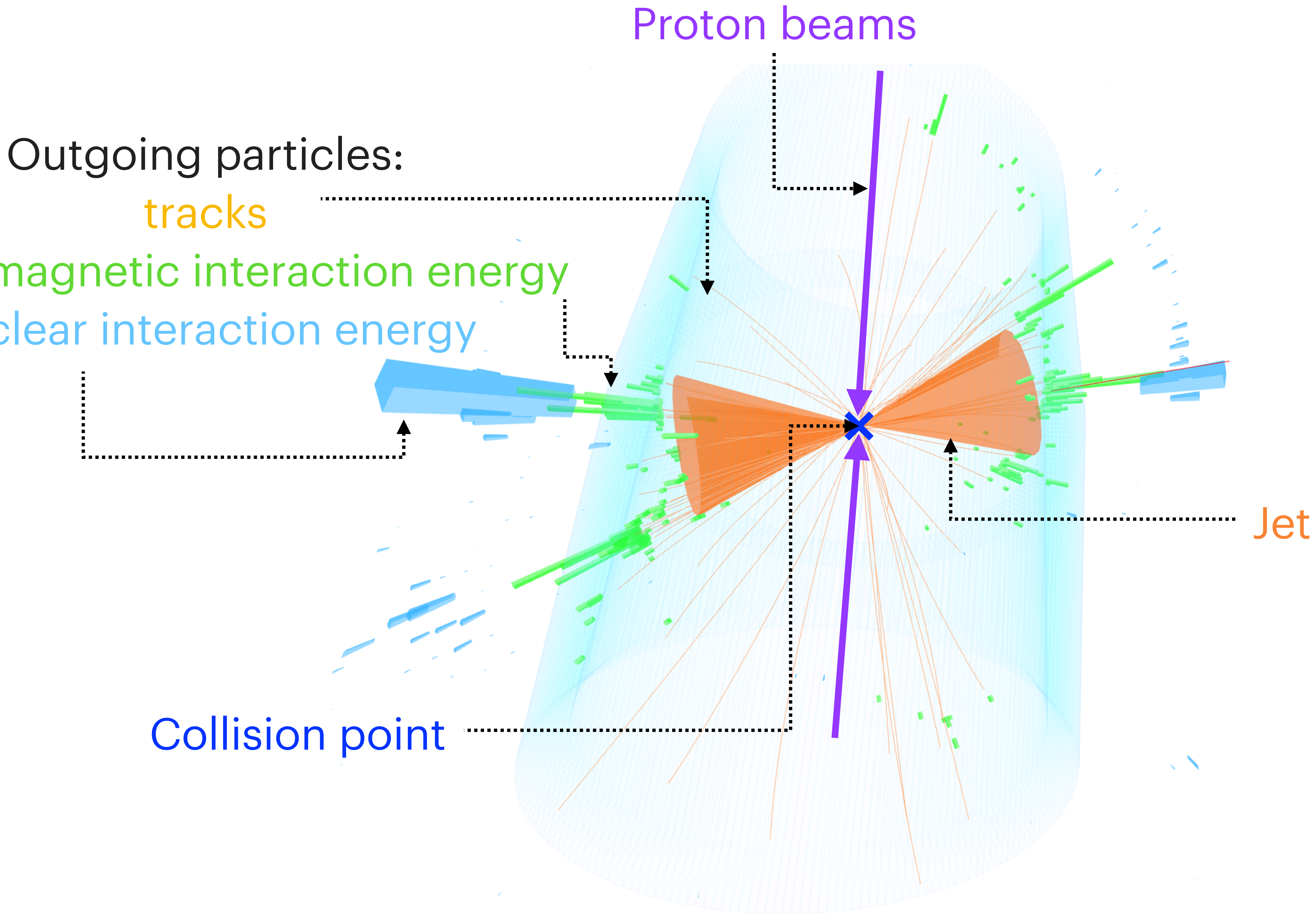
analyze ~1000 collisions / second

COMPACT MUON SOLENOID

- Specialized components to measure different particles
- 100 million channels

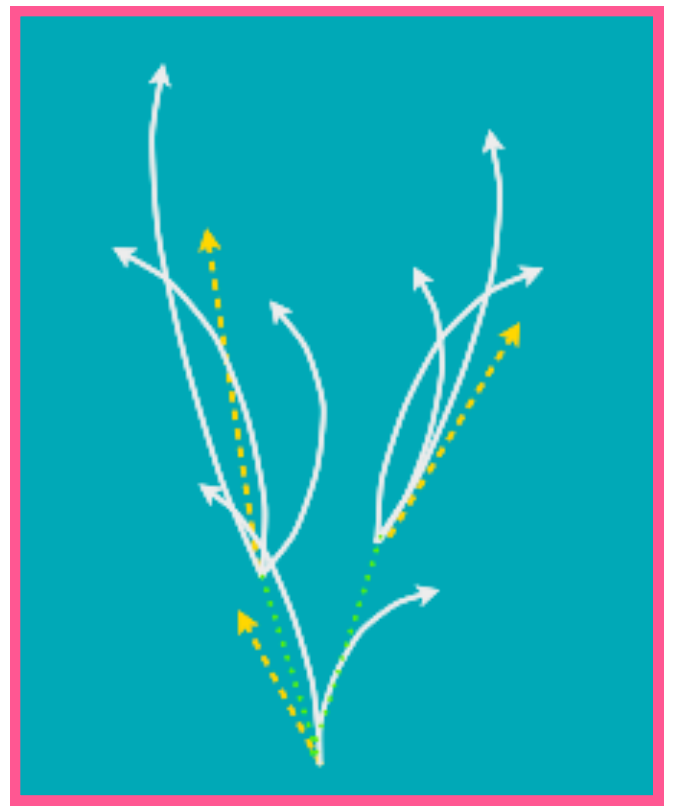


LHC RAW DATA TO "PARTICLE" CLOUDS



Collision event

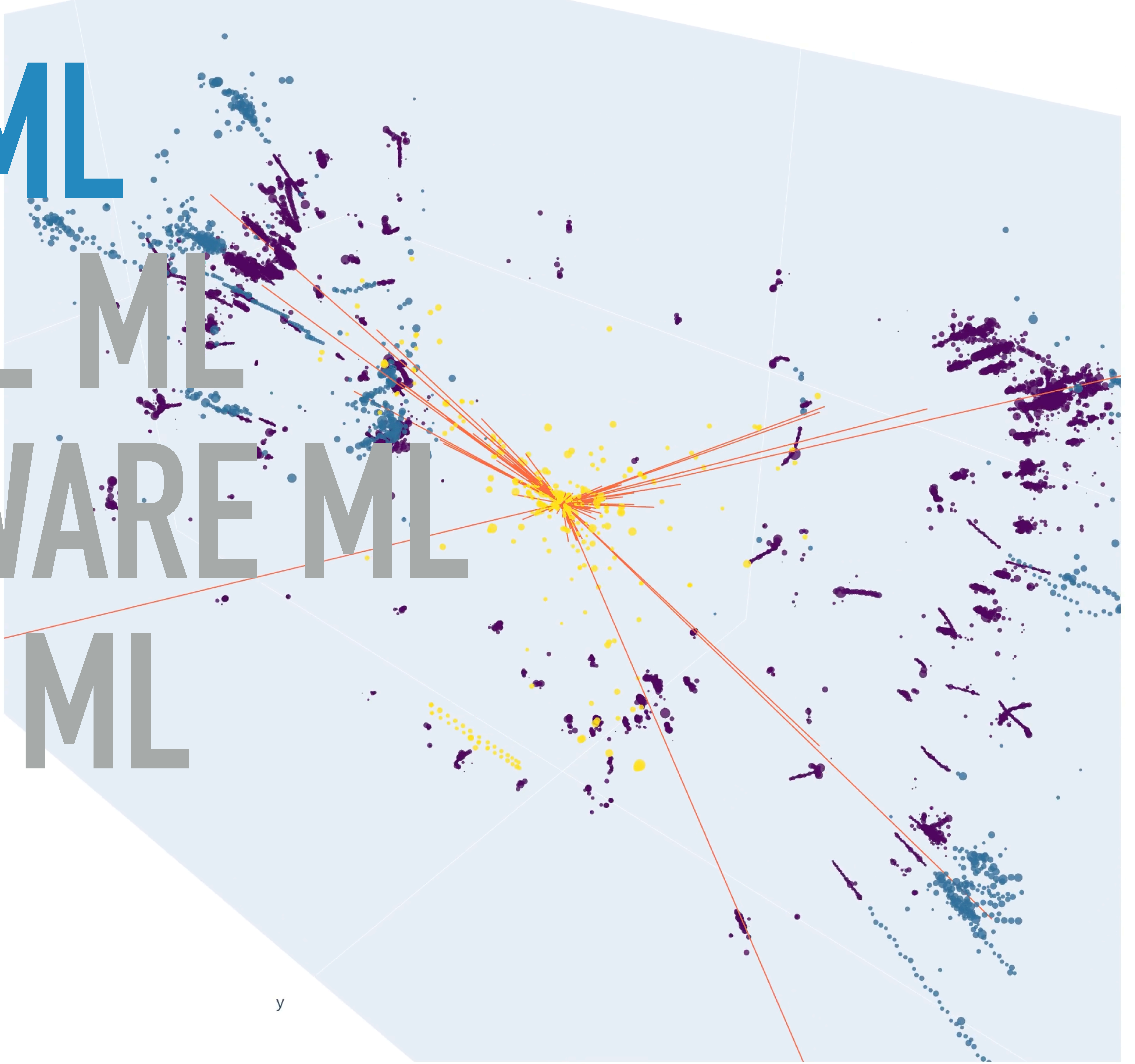
Higgs boson?



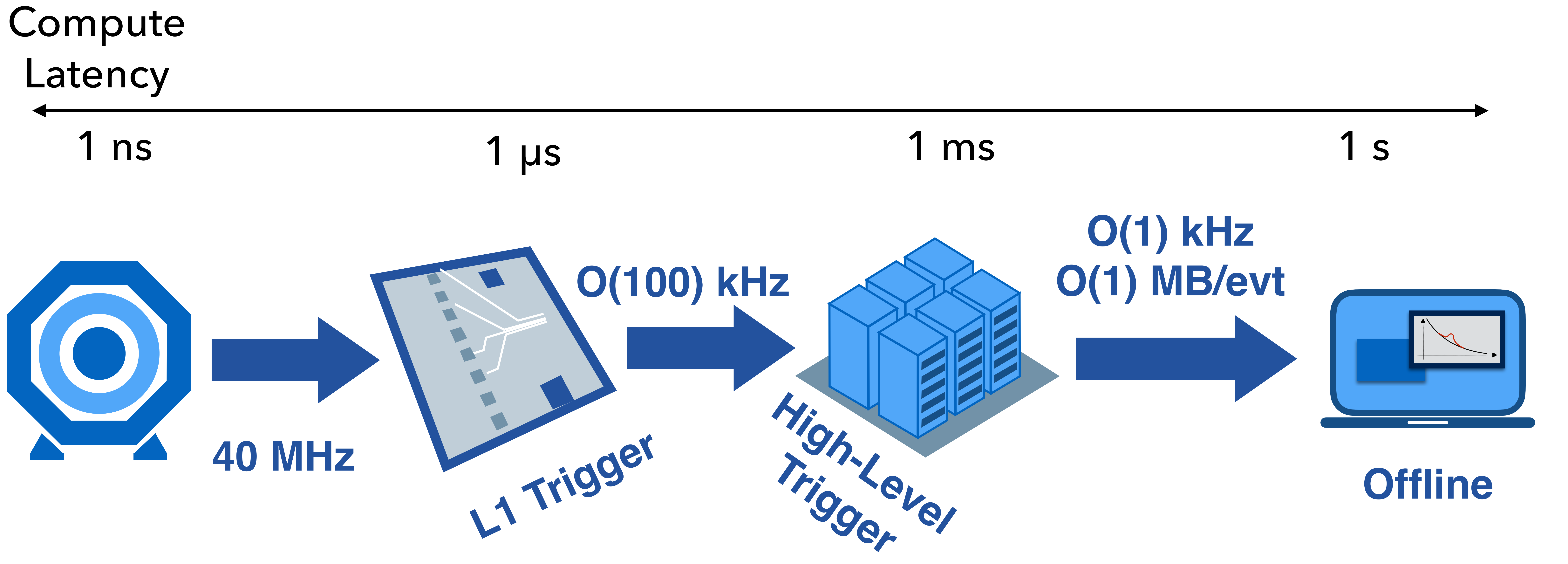
OUTLINE

- Machine learning has changed the way we do particle physics
  - It is an essential and versatile tool that we use to improve existing approaches
  - It enables fundamentally new approaches
- Highlight a few active areas of R&D, with public datasets and opportunities to collaborate!

ULTRAFAST ML
MULTIMODAL ML
PHYSICS-AWARE ML
GENERATIVE ML
OUTLOOK



LHC EVENT PROCESSING



ASICs	FPGAs	CPUs	CPUs
		GPUs	GPUs
			FPGAs
			Exabyte-scale datasets

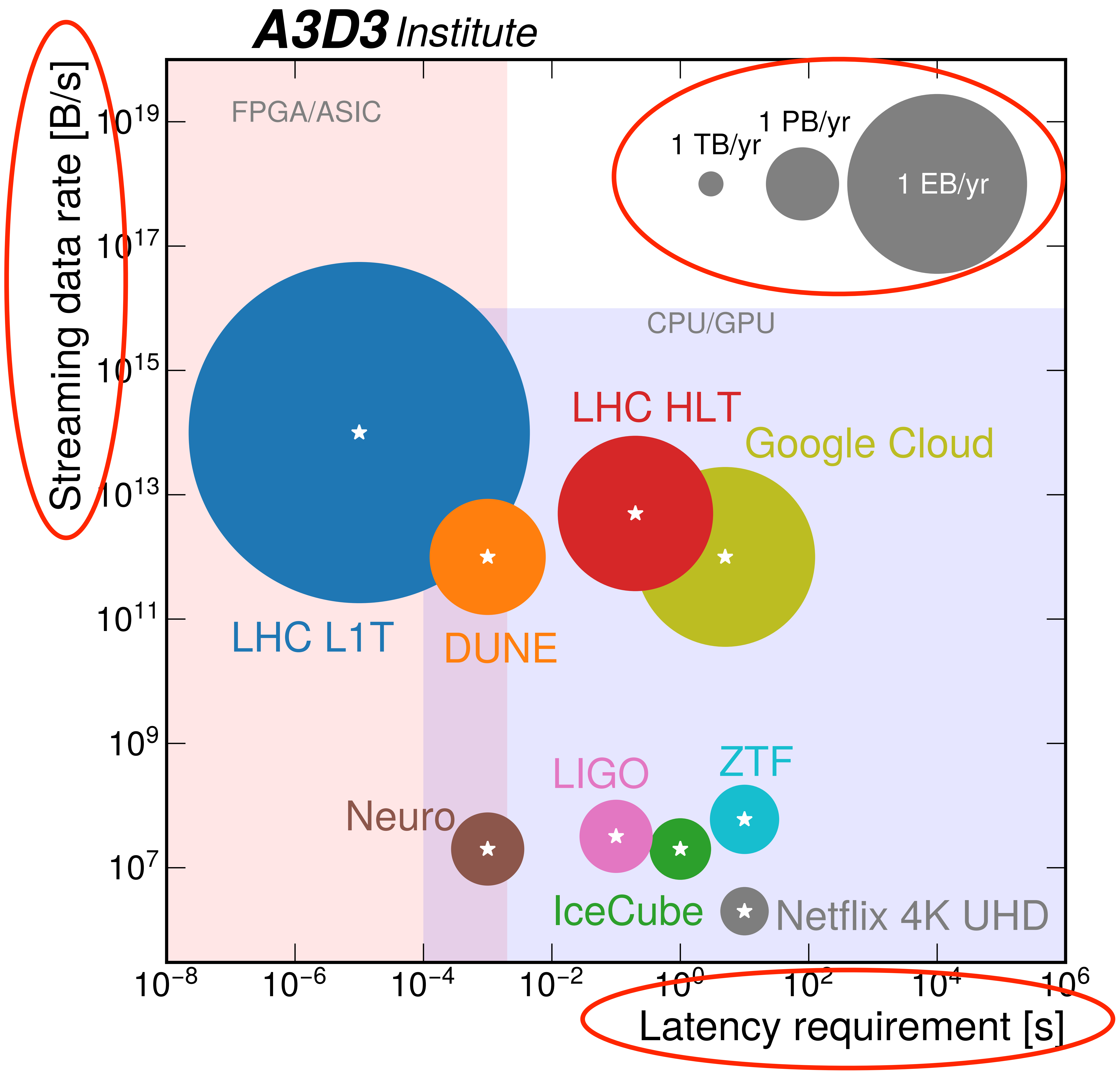
Challenges:

Each collision produces $O(10^3)$ particles

The detectors have $O(10^8)$ sensors

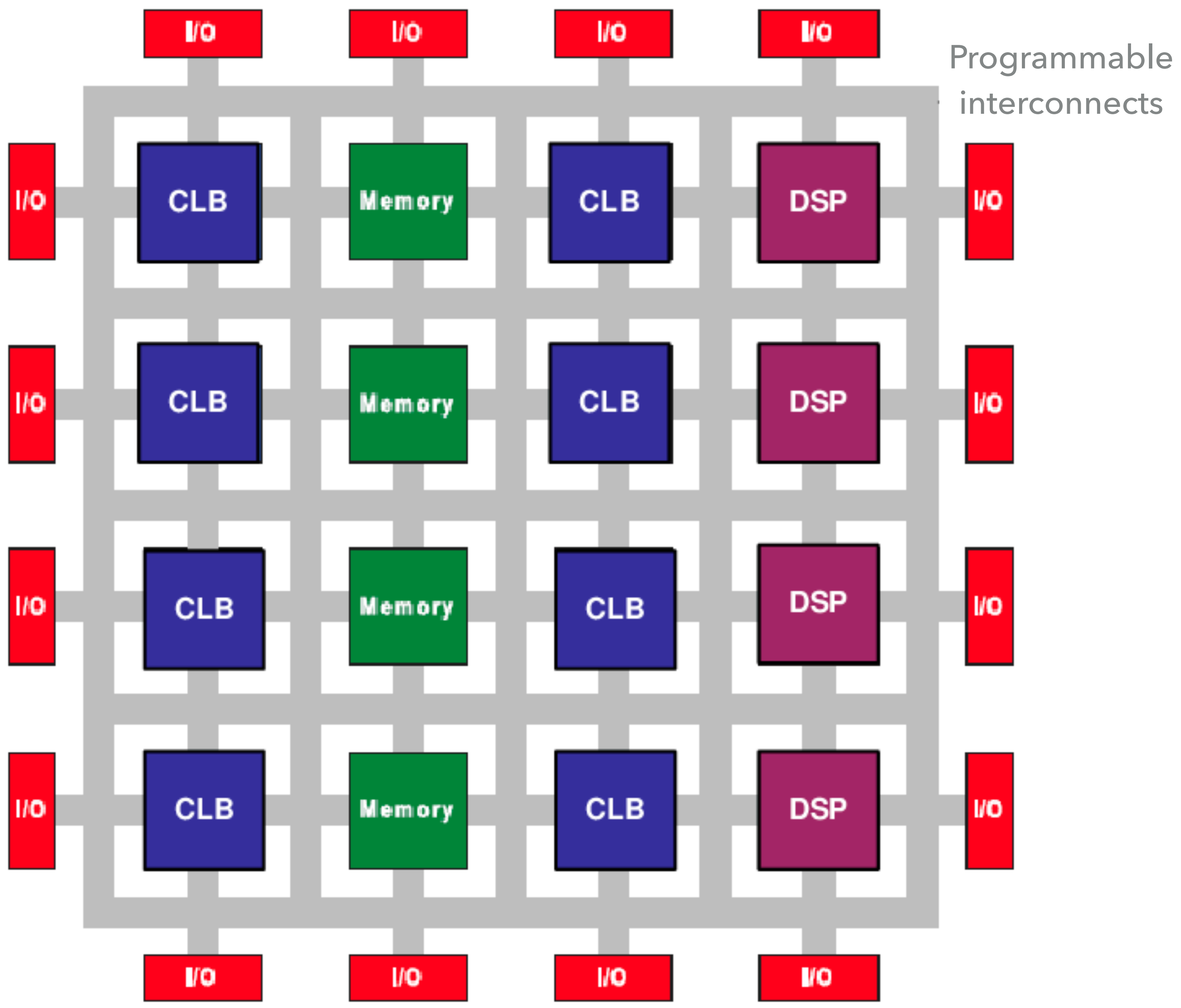
Extreme data rates of O(100 TB/s)

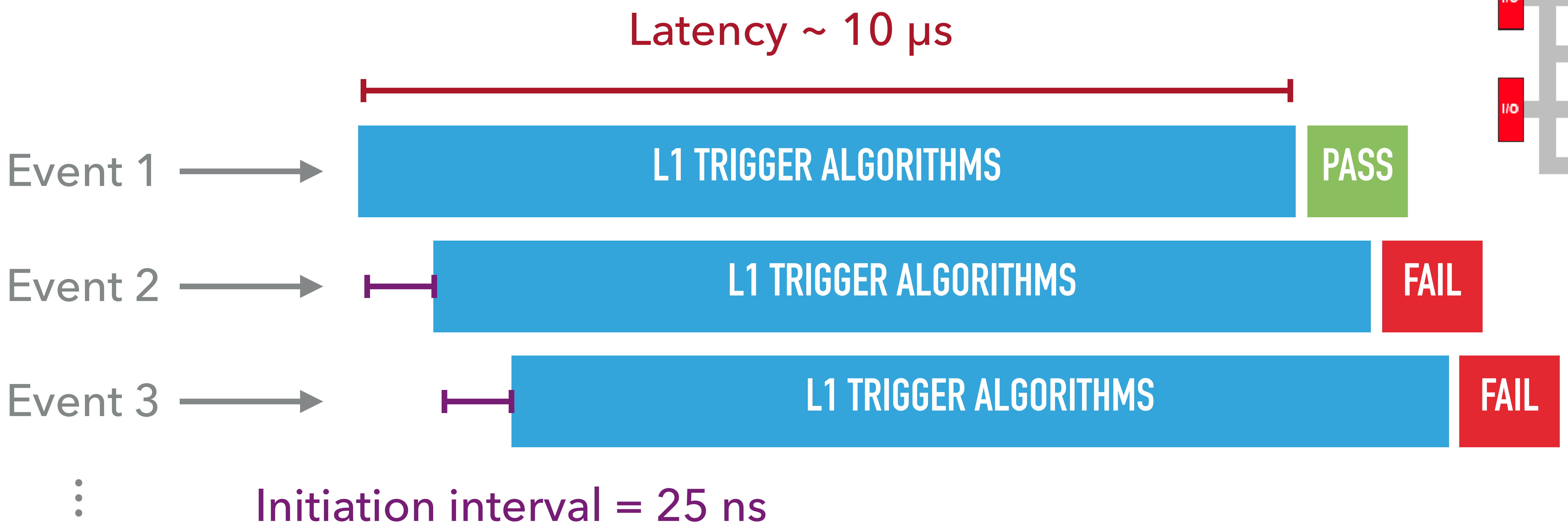
SCIENTIFIC ML CHALLENGES



WHAT MAKES THIS HARD?

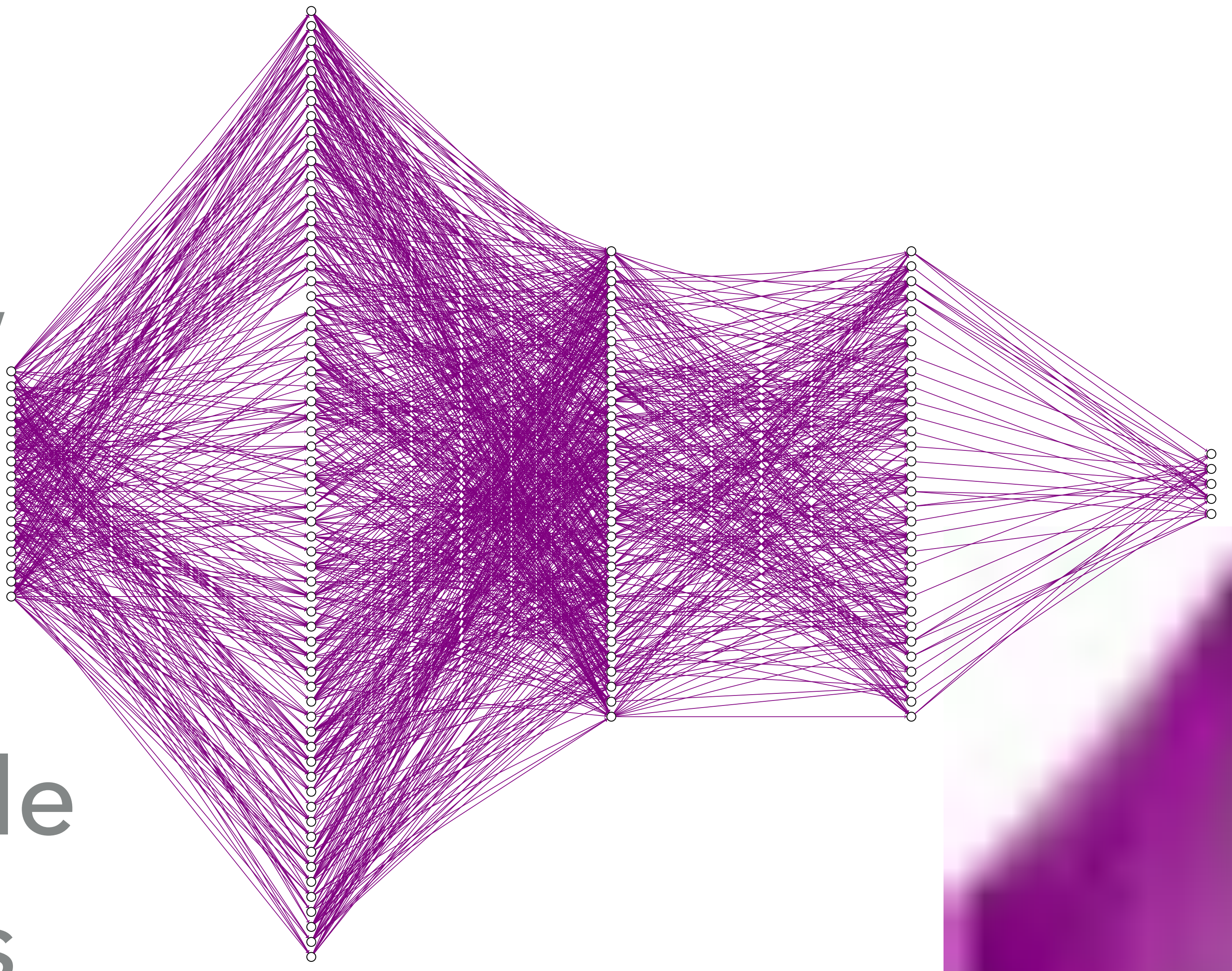
- Reconstruct all events and reject 98% of them in O(10) μs
  - Algorithms have to be <1 μs and process new events every 25 ns
- Latency necessitates all **FPGA** design
  - Algorithms have to fit on <1 FPGA
- How can we satisfy these constraints?

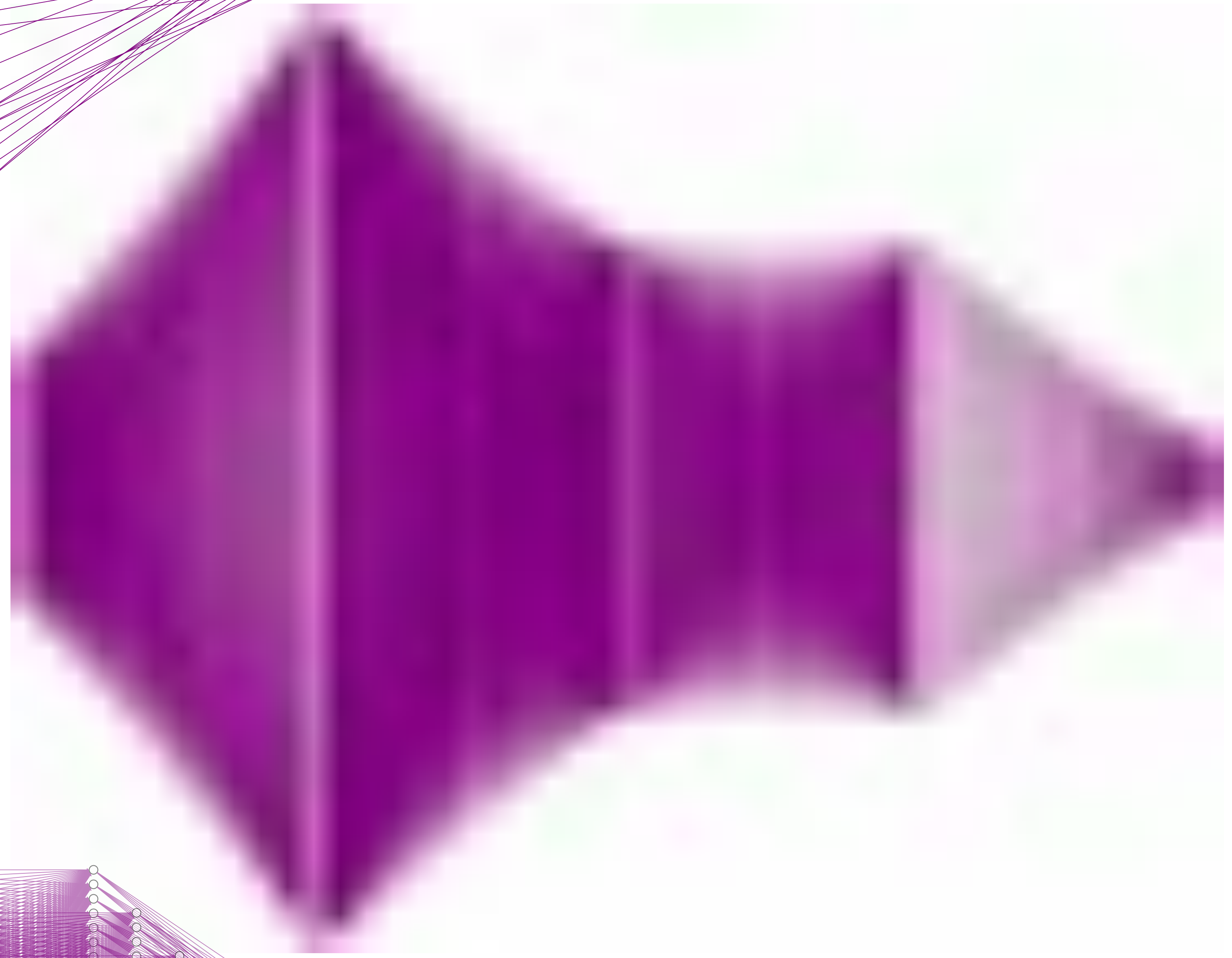


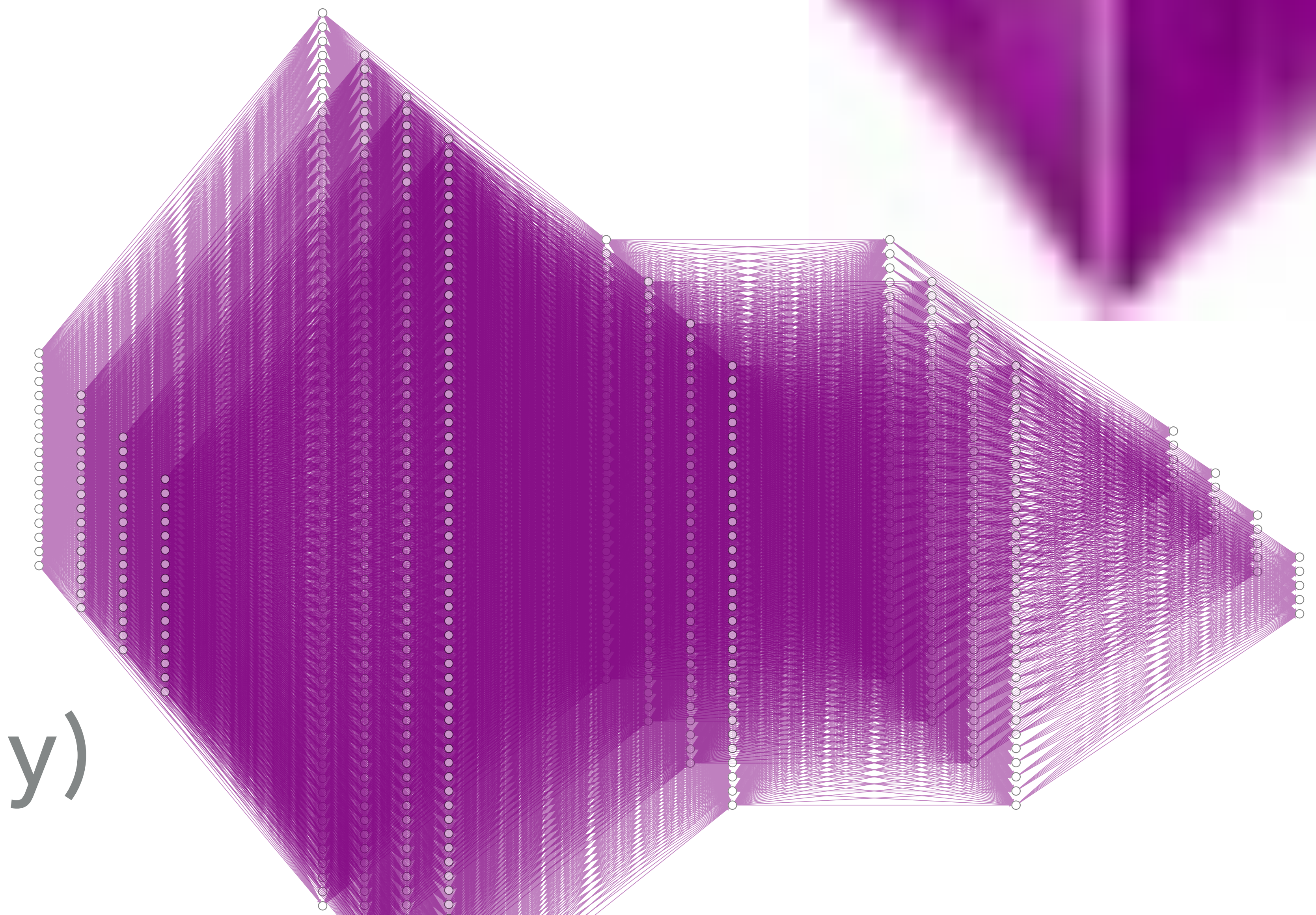


CODESIGN

- **Codesign**: intrinsic development loop between ML design, training, and implementation
- Pruning
  - Maintain high performance while removing redundant operations
- Quantization
  - Reduce precision from 32-bit floating point to 16-bit, 8-bit, …
- Parallelization
  - Balance parallelization (how fast) with resources needed (how costly)



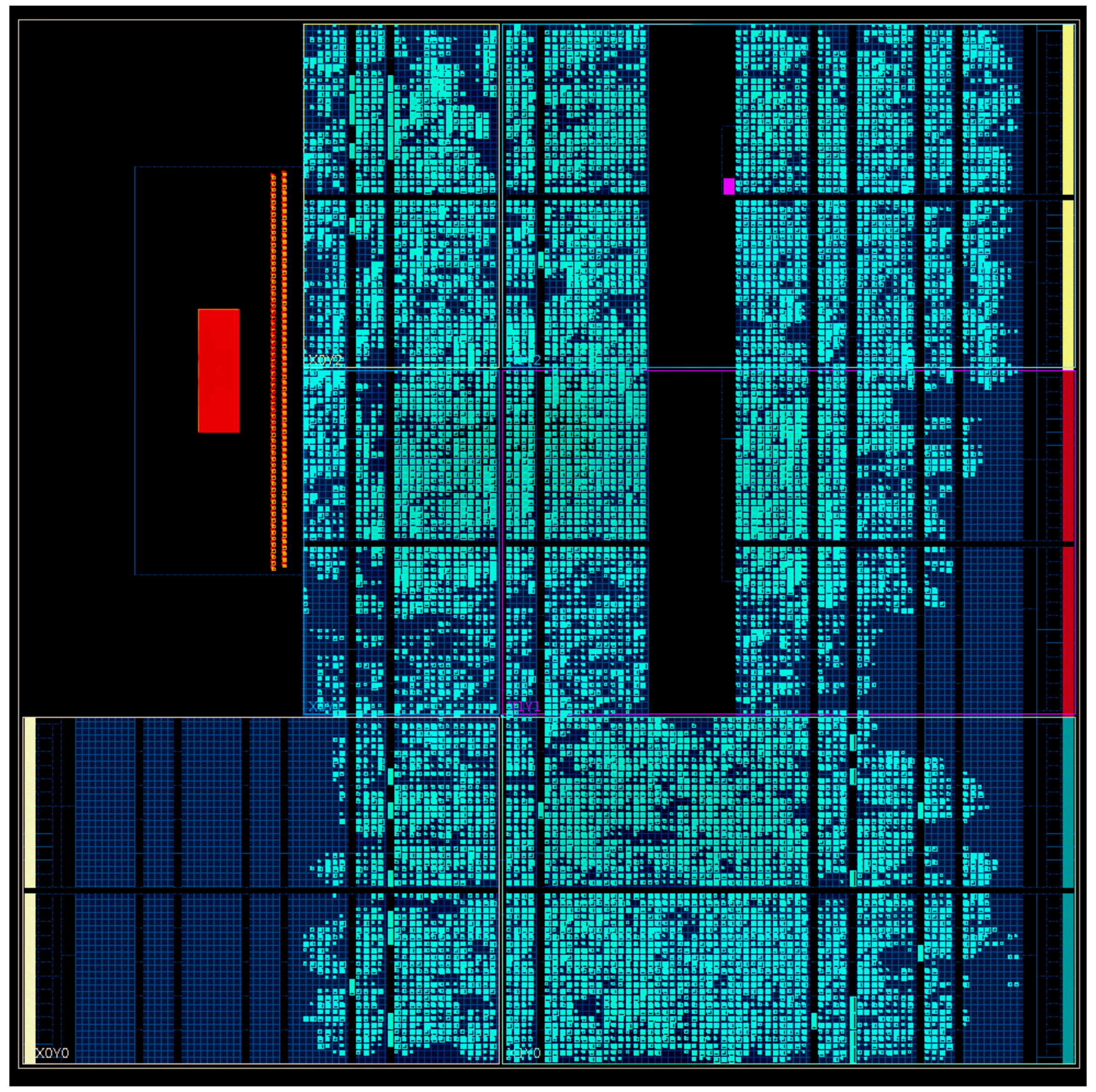


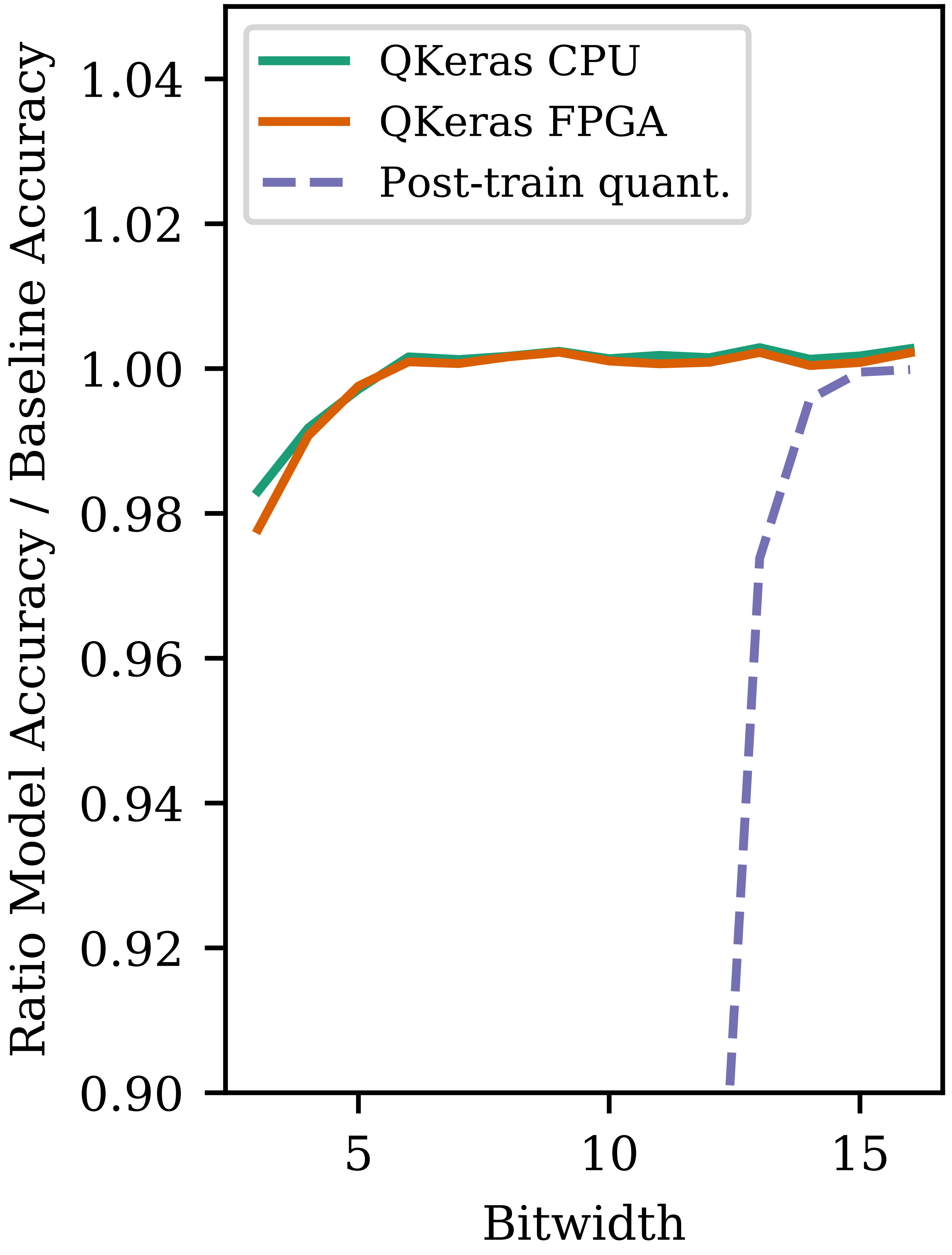


QUANTIZATION-AWARE TRAINING: RESULTS

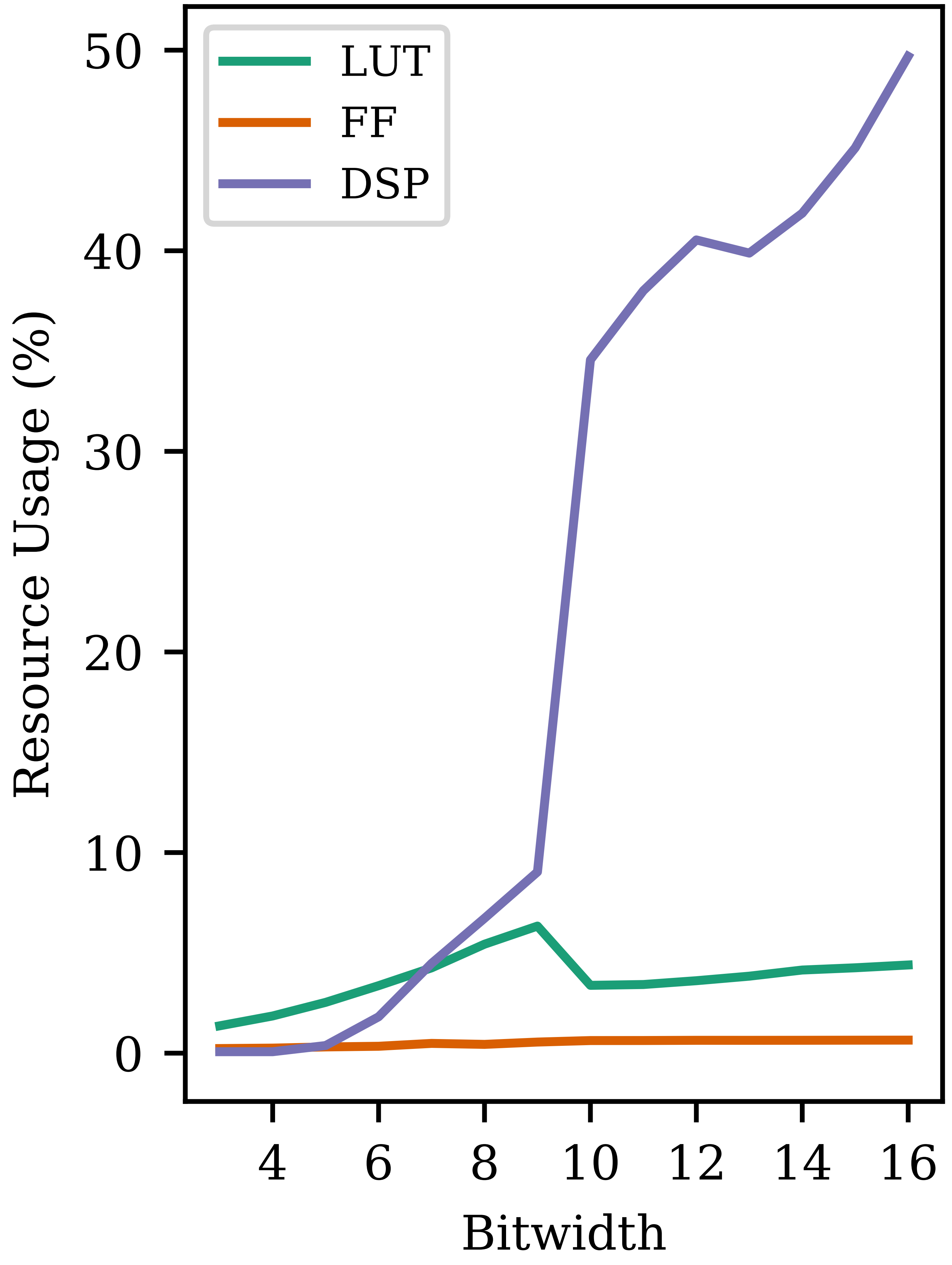
Nat. Mach. Intell. 3, 675 (2021)

- Small NN benchmark correctly identifies particle "jets" 70-80% of the time
- Full performance with 6 bits instead of 14 bits
- Much smaller fraction of resources



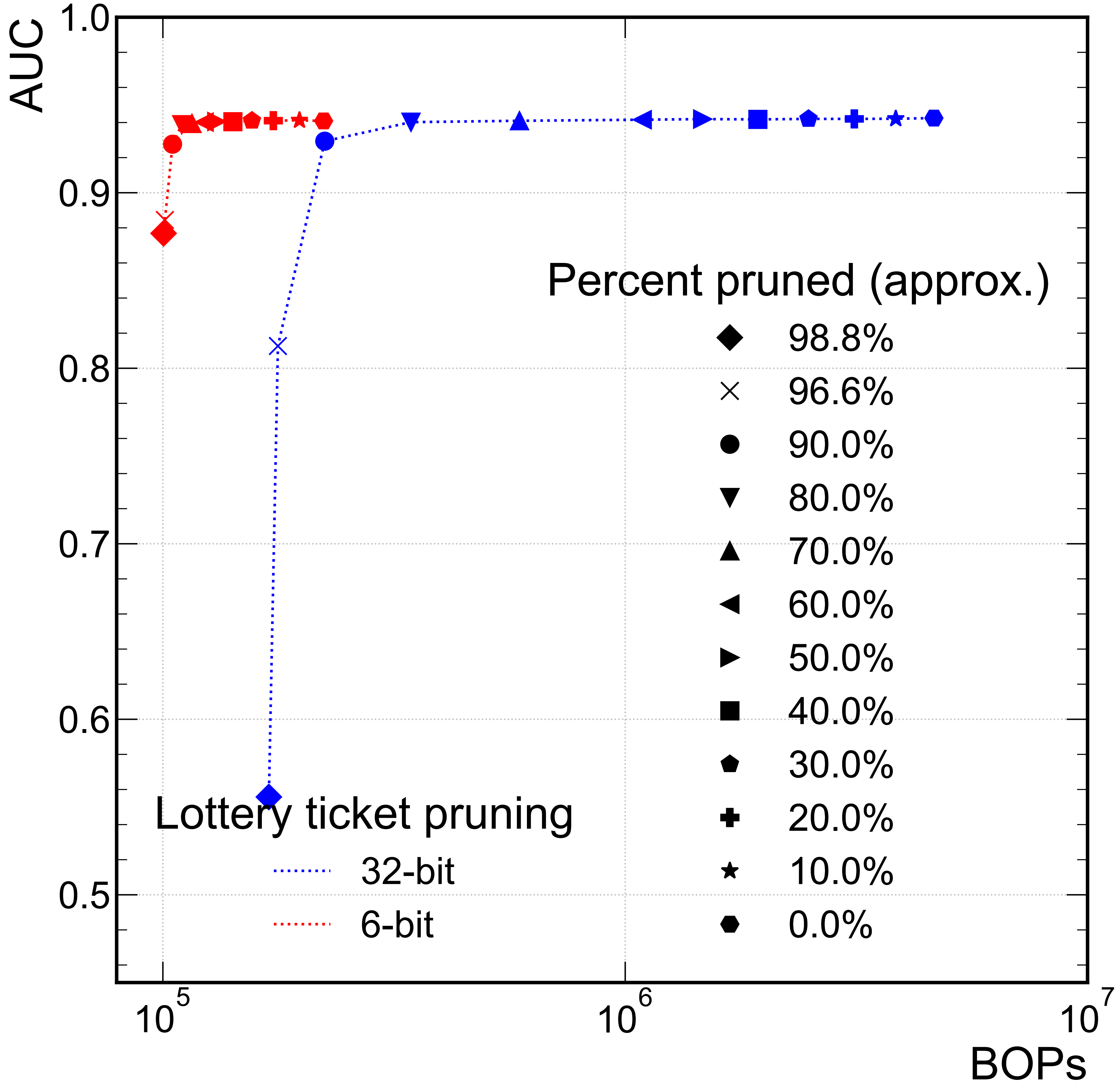


Xilinx VU9P



PRUNING + QUANTIZATION-AWARE TRAINING

- Quantization-aware pruning (QAP): iterative pruning further reduces hardware complexity of a quantized model
- After QAP, $50\times$ reduction in bit operations compared to the 32-bit, unpruned model



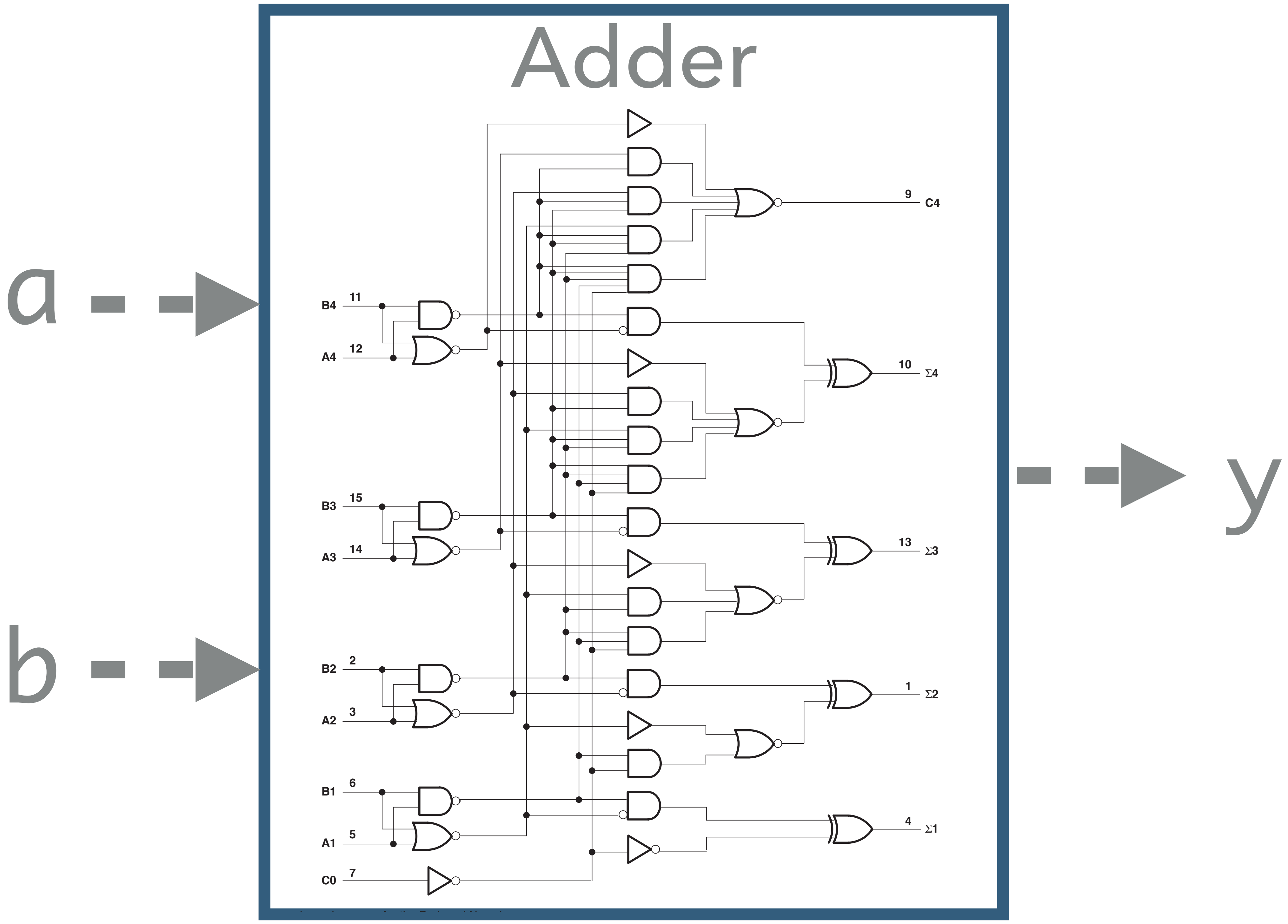
Bit operations (BOPs) definition: arXiv:1804.10969

PROGRAMMING HARDWARE (FPGAS)

- Say you want to program an "adder" function on an FPGA

```
module adder(
        input  wire [4:0] a,
        input  wire [4:0] b,
        output wire [4:0] y
);
        assign y = a + b;

endmodule
```

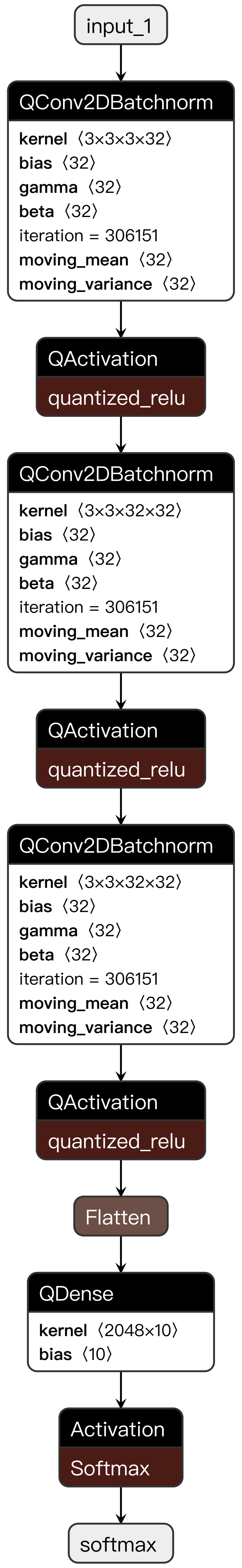


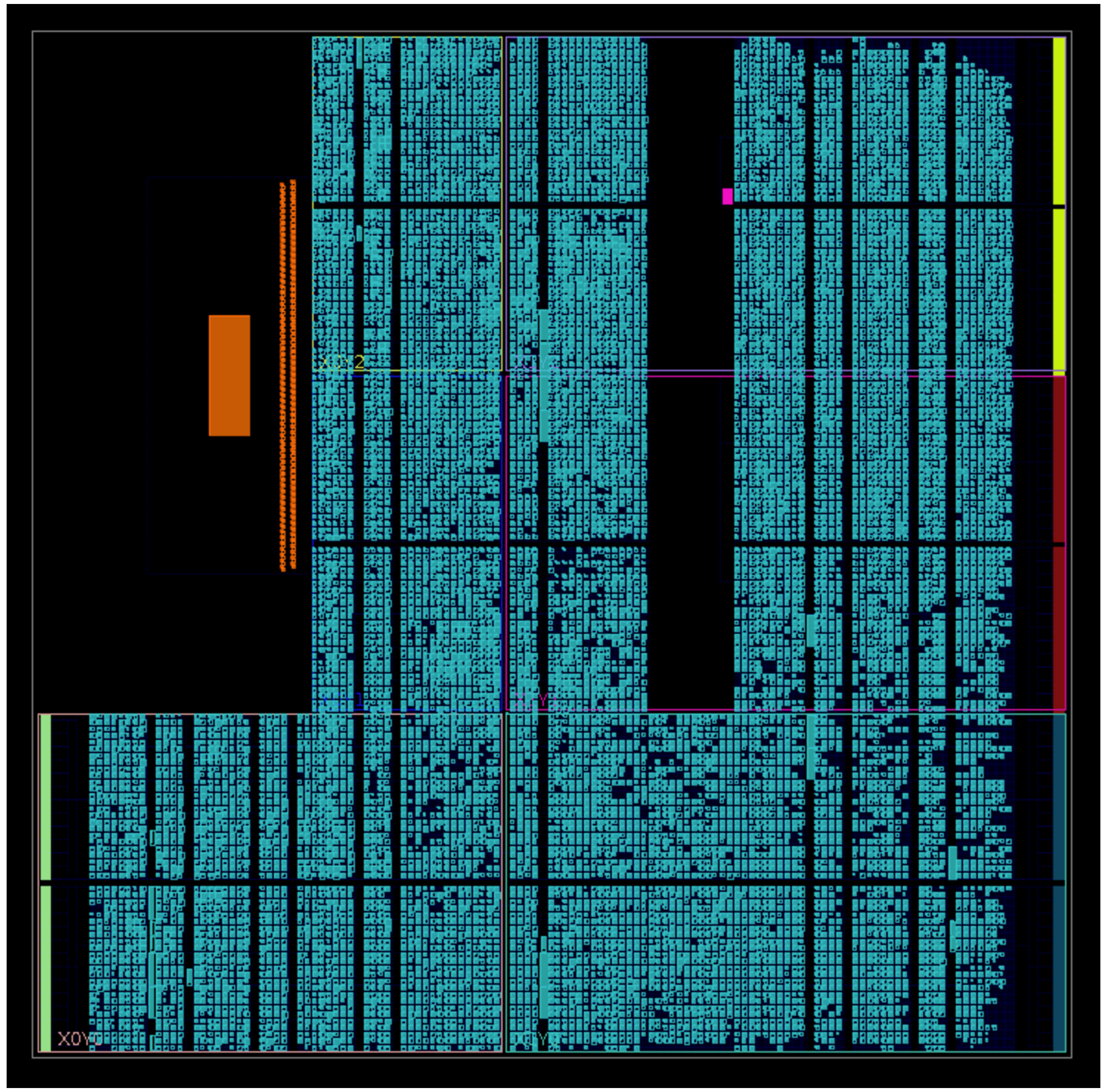
Synthesis

- Register transfer-level (RTL) code is "synthesized" into gates

PROGRAMMING HARDWARE (FPGAS)

- What if instead we specify an AI model (e.g., in QONNX)

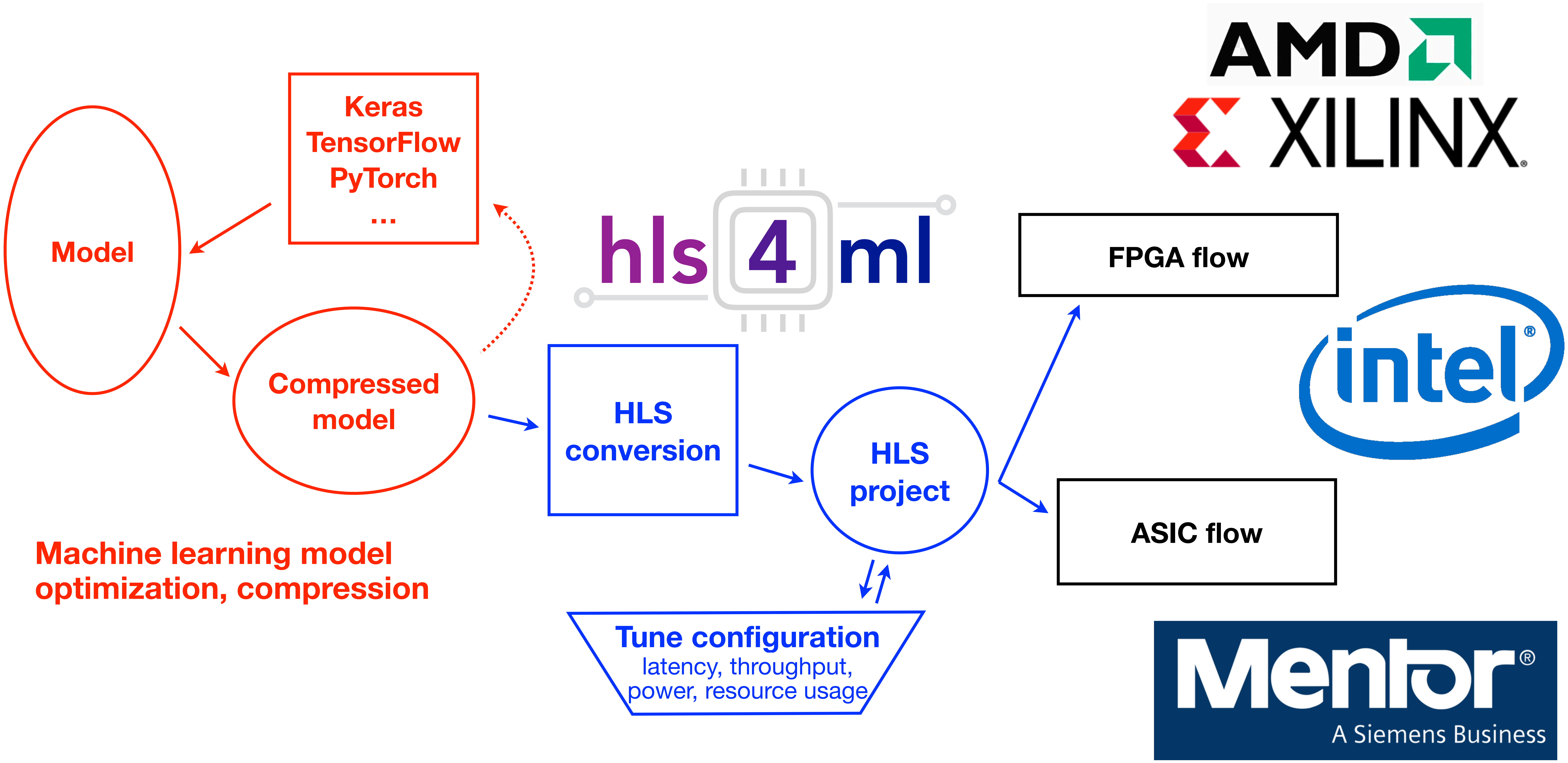




High-Level Synthesis

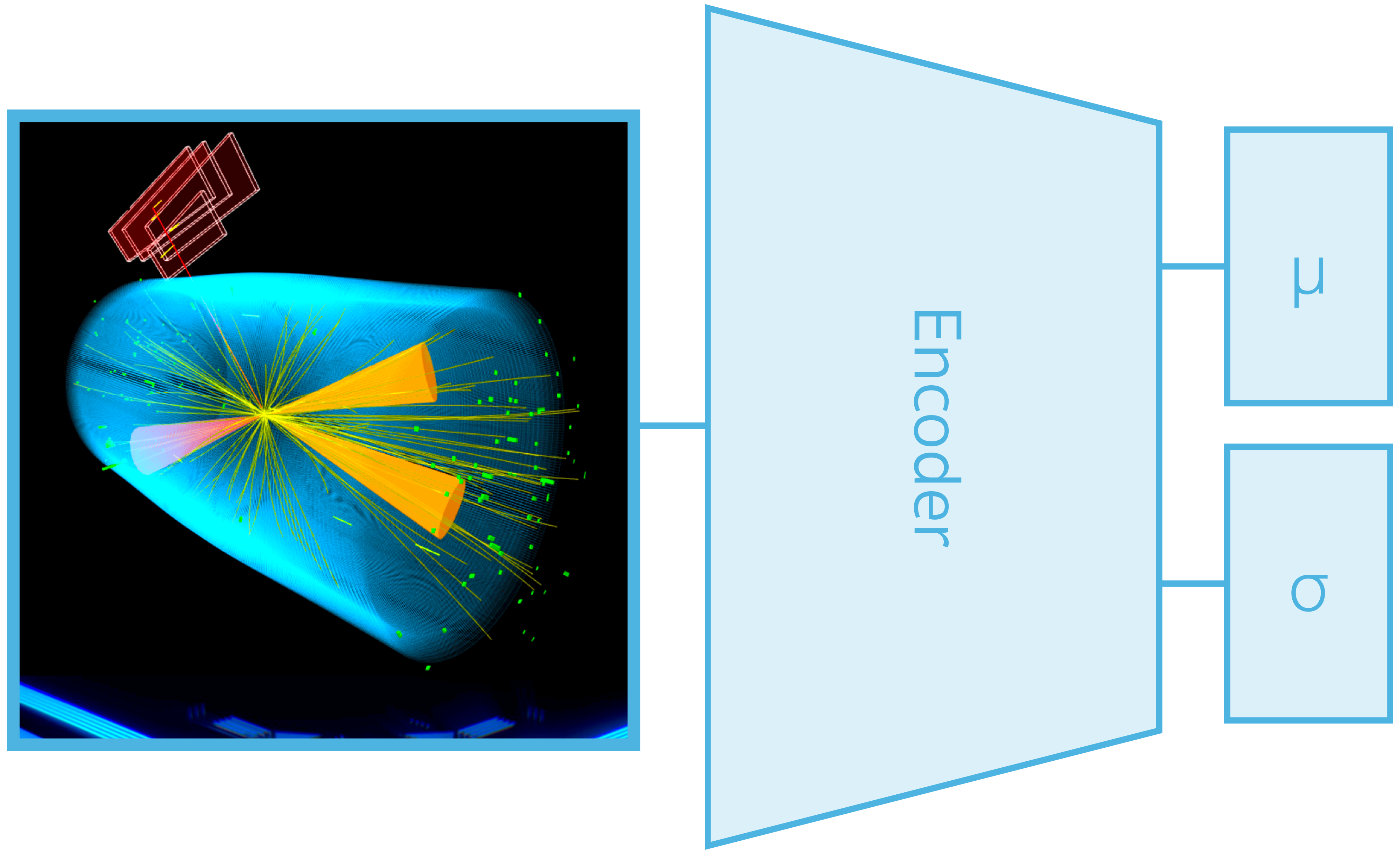
DESIGN EXPLORATION WITH HLS4ML

- hls4ml for scientists or ML experts to translate ML algorithms into RTL firmware



APPLICATION: ANOMALY DETECTION AT 40 MHZ

- Challenge: if new physics has an unexpected signature that doesn't align with existing triggers, precious signal events will be discarded
- Can we use unsupervised algorithms to detect non-SM-like anomalies?
  - Autoencoders (AEs): compress input to a smaller dimensional latent space then decompress and calculate difference
  - Variational autoencoders (VAEs): model the latent space as a probability distribution; possible to detect anomalies purely with latent space variables

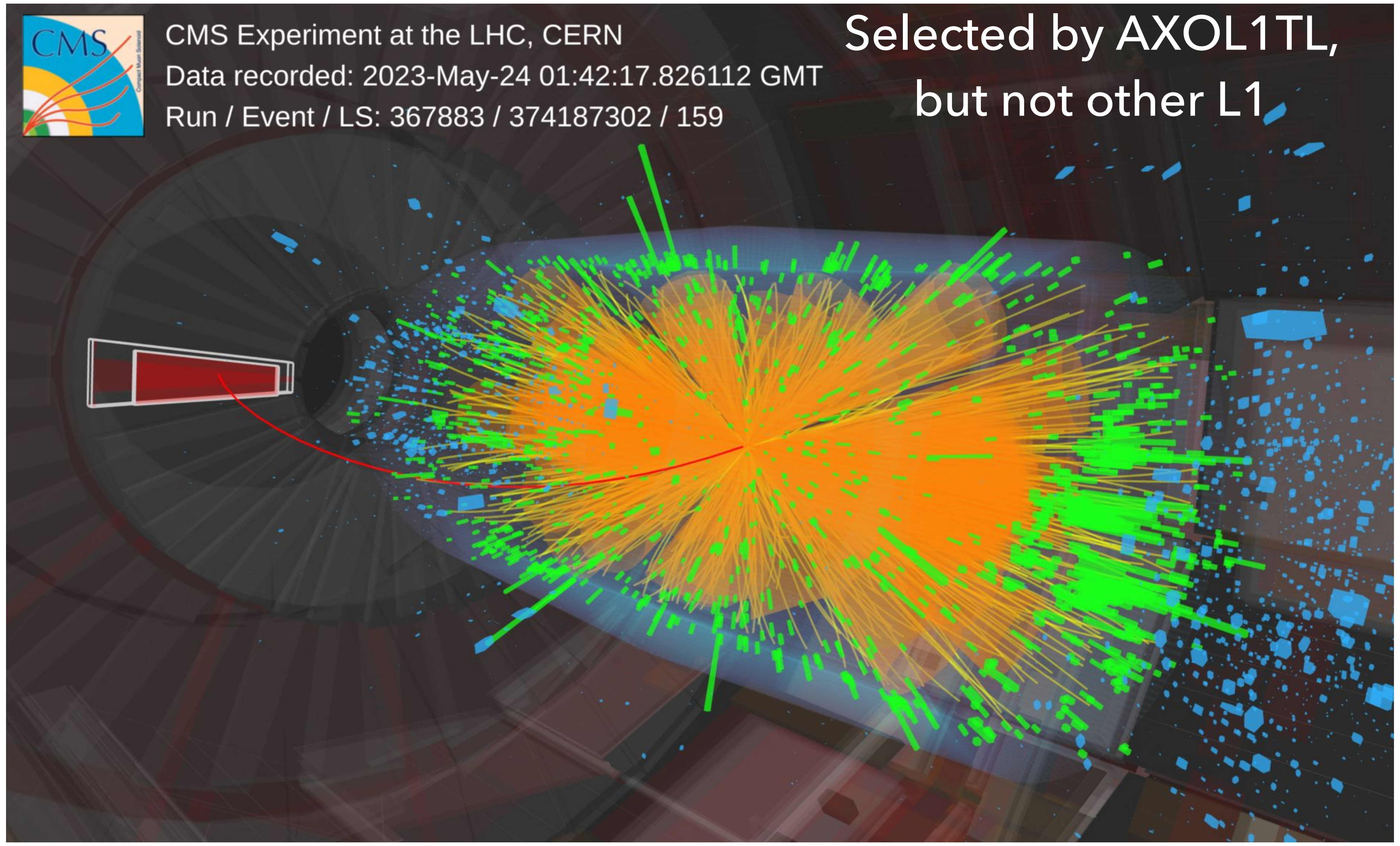


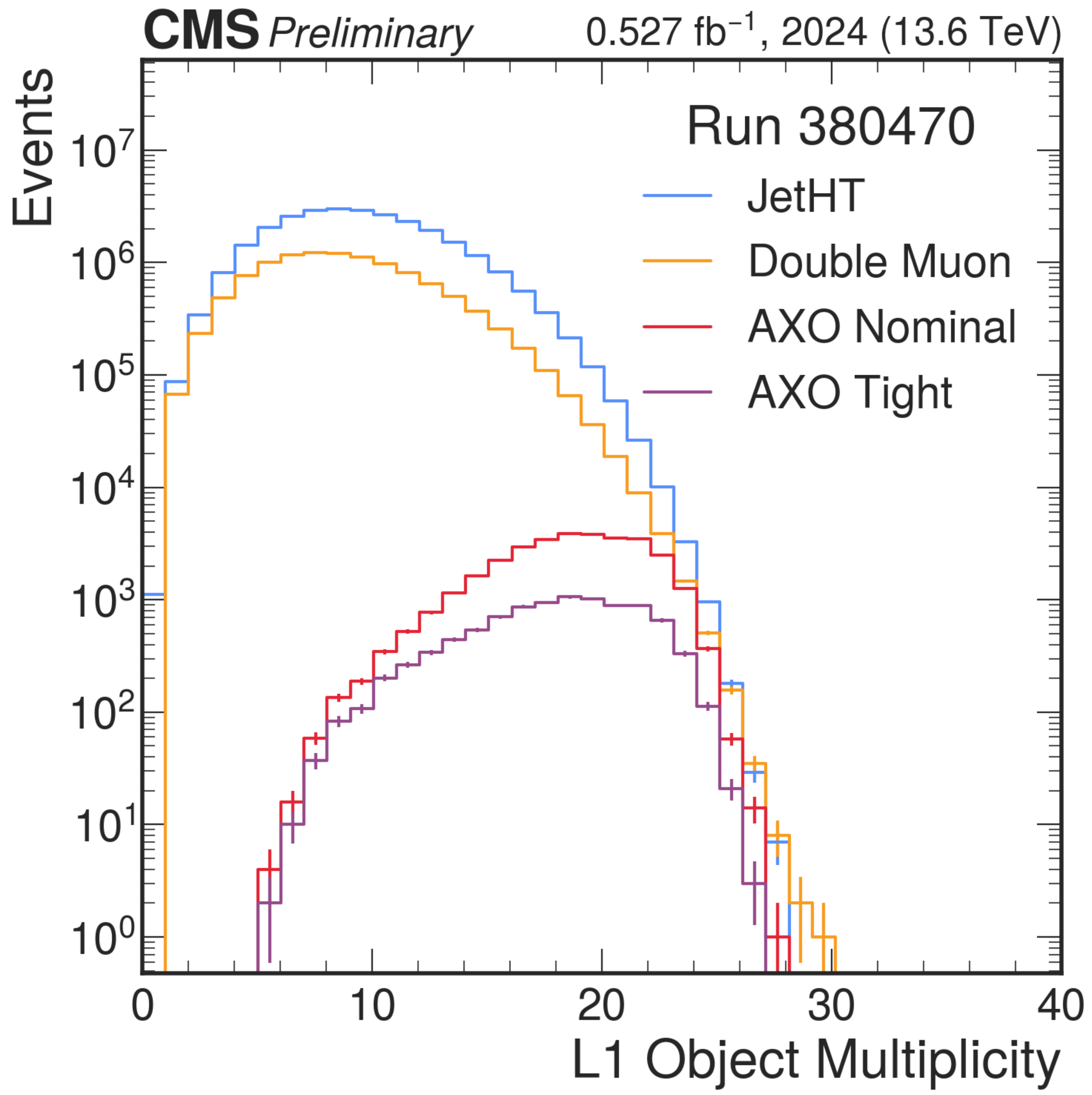
Key observation: Can build an anomaly score from the latent space of VAE directly! No need to run decoder!

$$R_z = \sum_i \frac{\mu_i^2}{\sigma_i^2}$$

ANOMALY DETECTION @ LEVEL-1 TRIGGER

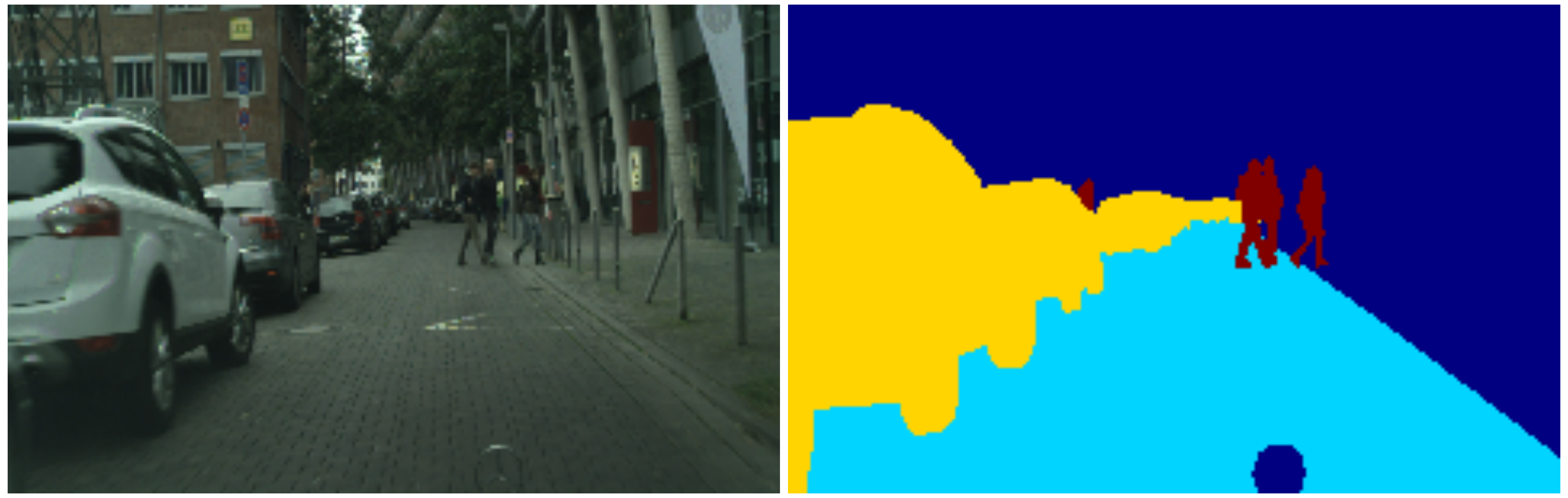
- AXOL1TL anomaly detection algorithm for the trigger based on a variational autoencoder
- Selects unique events relative to existing triggers
- Preference for high multiplicity events

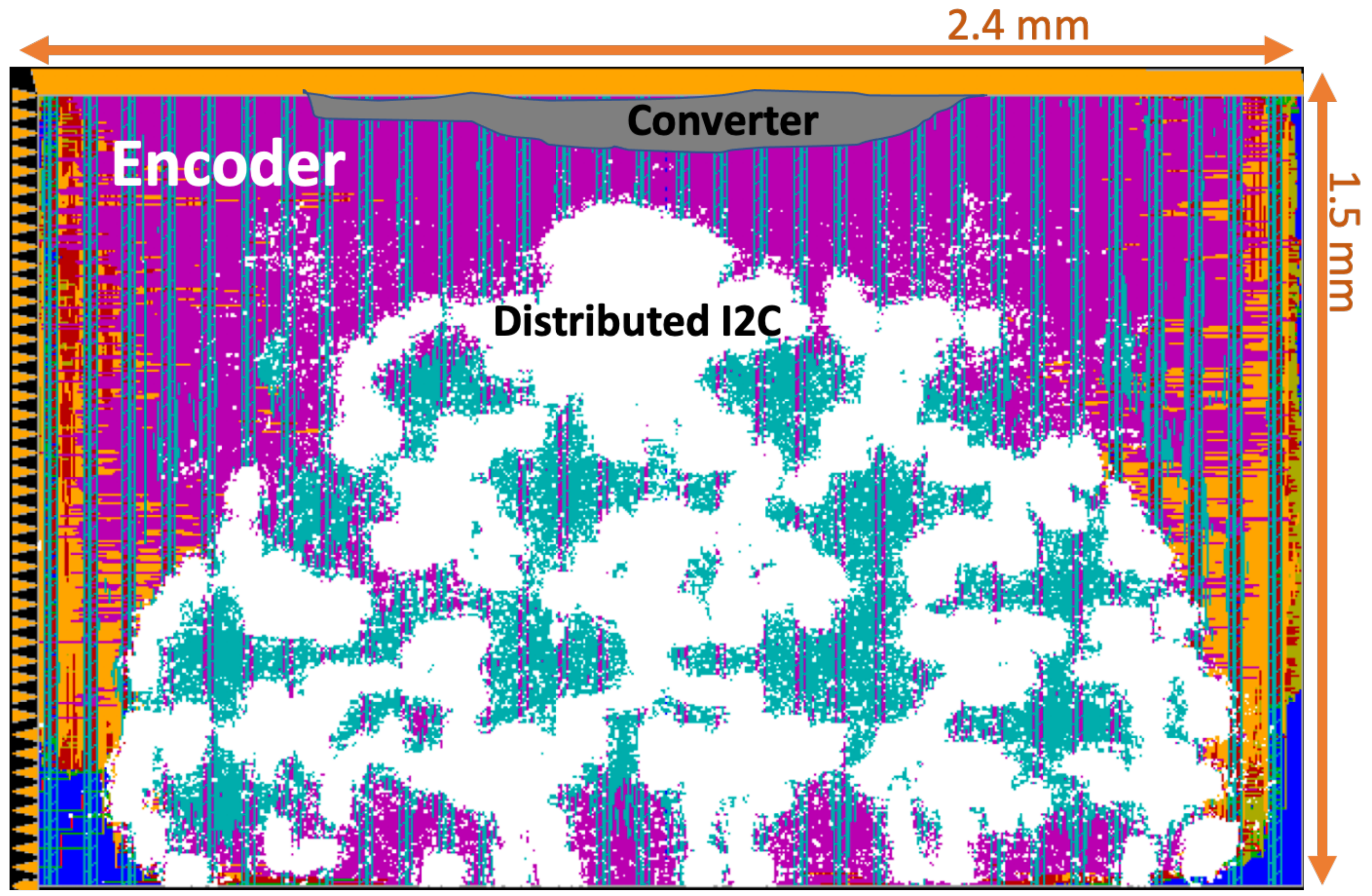


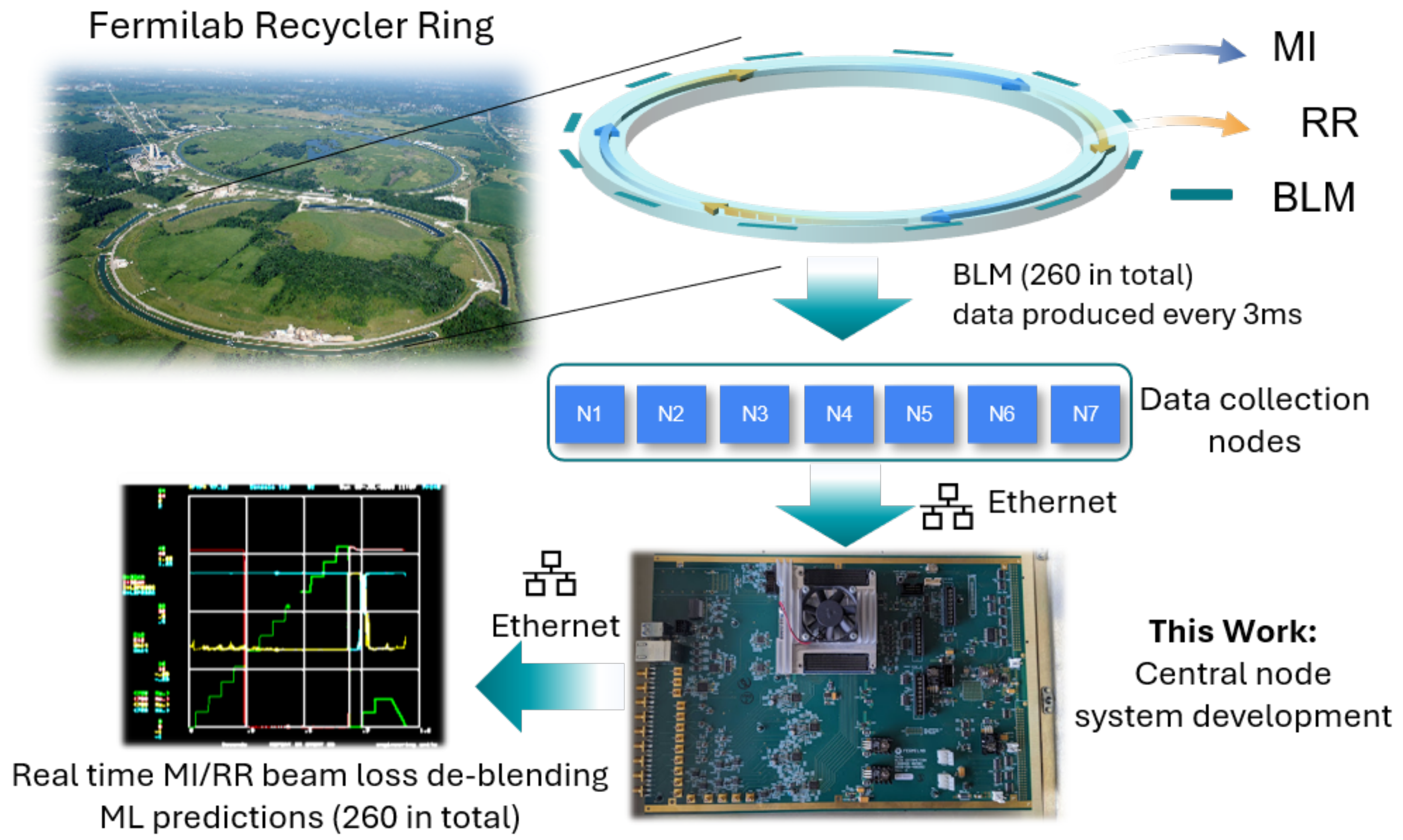


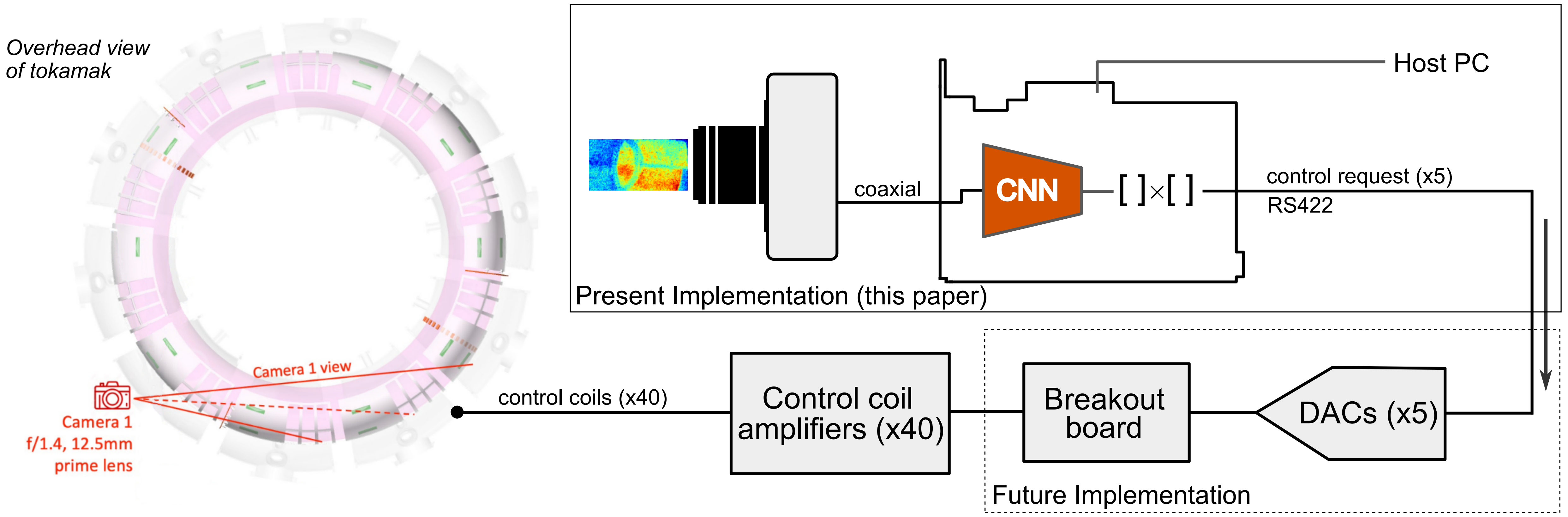
FAST ML APPLICATIONS BEYOND HEP

- Though hls4ml developed for particle physics, has seen widespread use for
  - Self-driving cars [arXiv:2205.07690]
  - Fusion devices [arXiv:2312.00128]
  - Steering particle beams [arXiv:2011.07371, arXiv:2311.05716]
  - Data compression at the edge [arXiv:2105.01683]





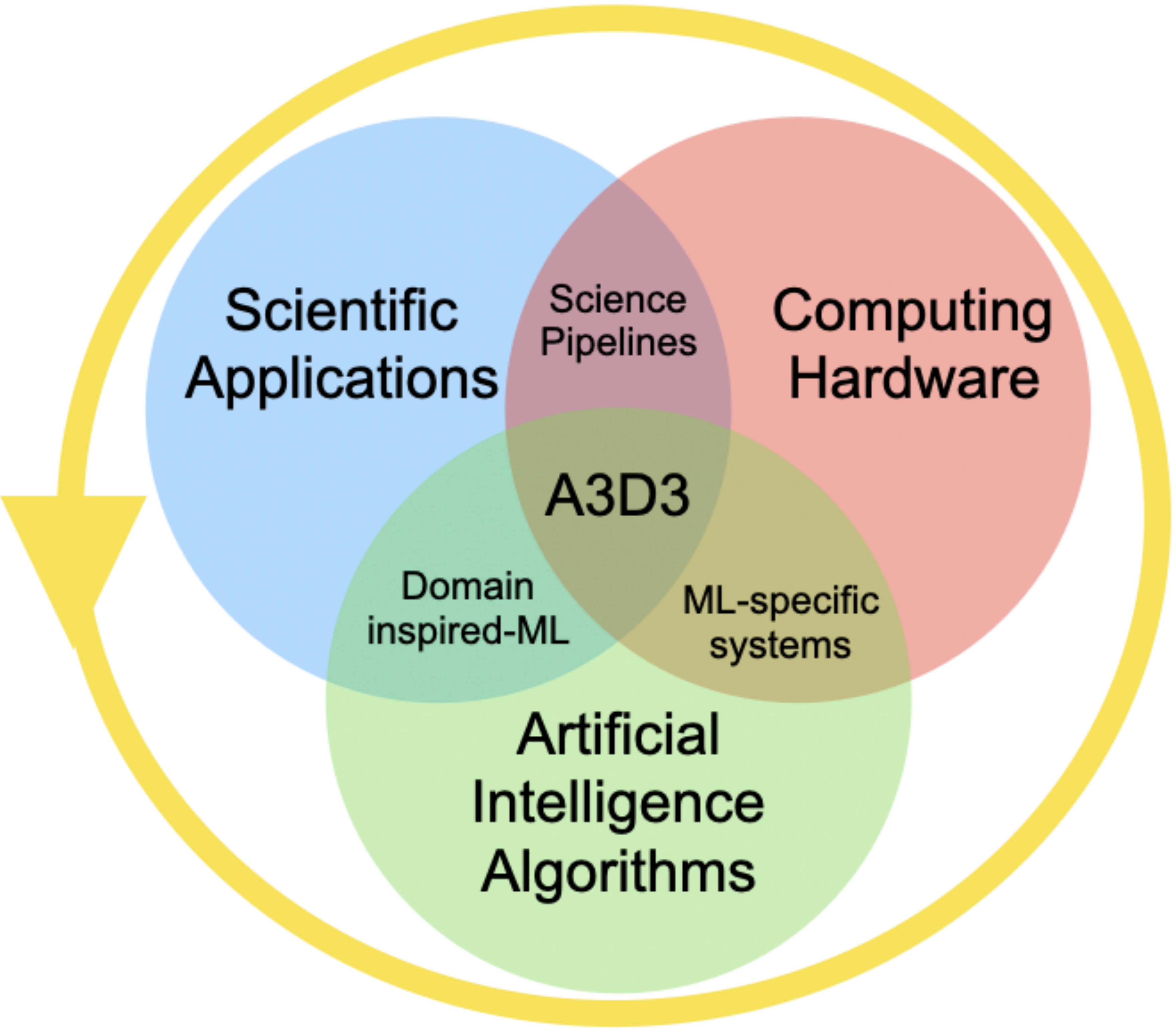




NSF INSTITUTE: A3D3

- Tightly coupled organization of domain scientists, computer scientists, and engineers that unite three core components which are essential to achieve real-time AI to transform science: AI techniques, Computing Hardware, Scientific Applications
- Check the [a3d3.ai](a3d3.ai) for events and more information!

PHY-2117997



Accelerated AI Algorithms for Data-Driven Discovery

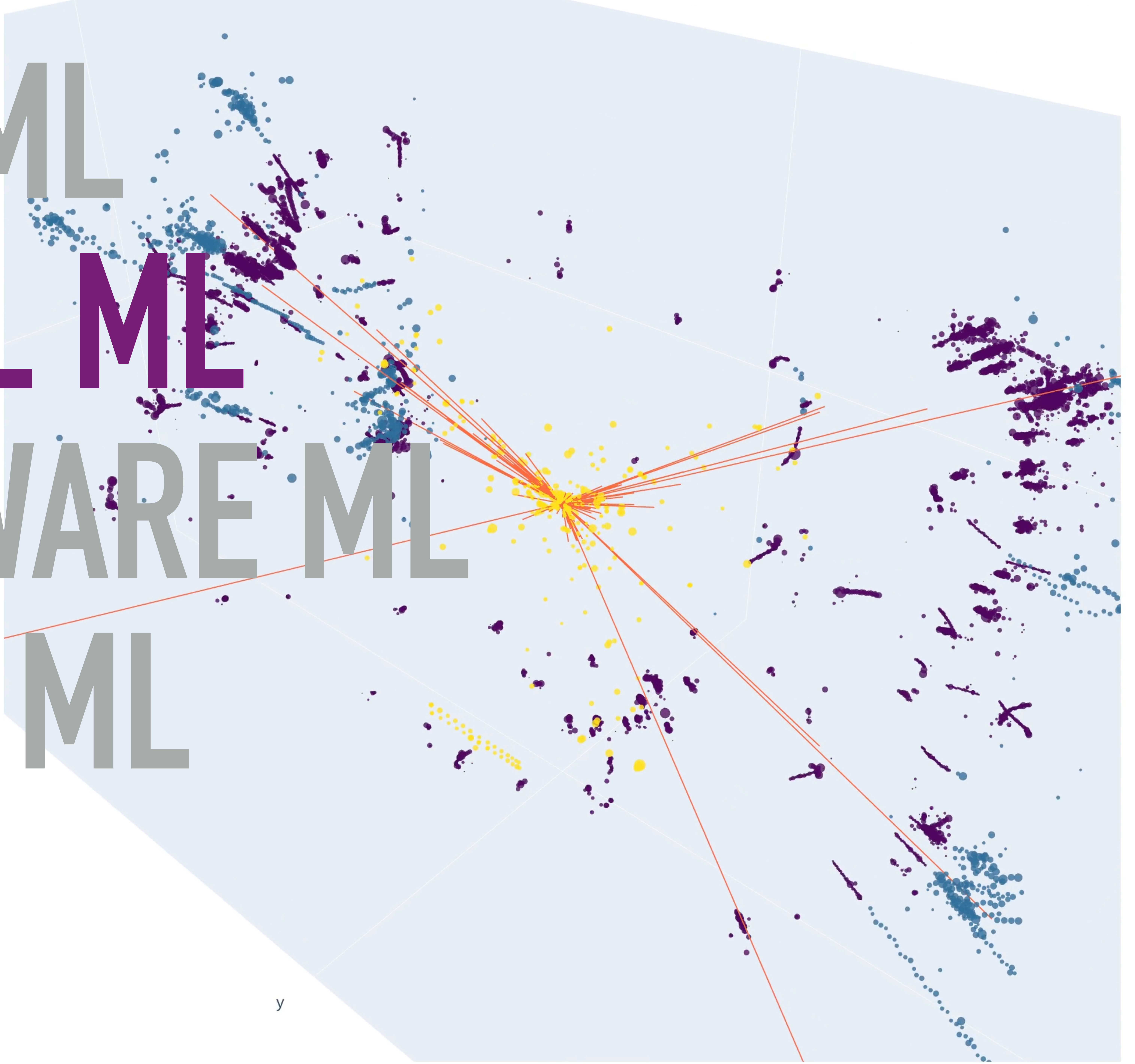
ULTRAFAST ML

MULTIMODAL ML

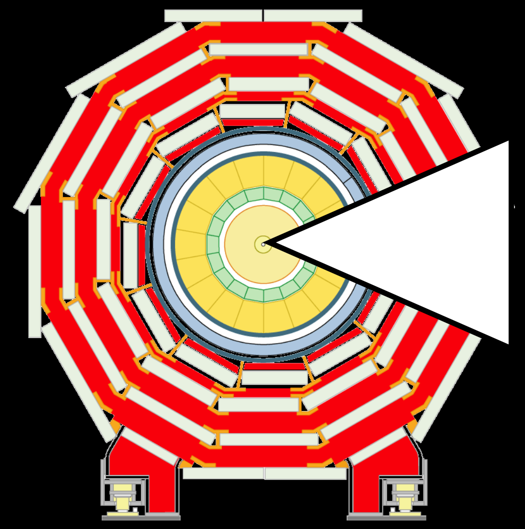
PHYSICS-AWARE ML

GENERATIVE ML

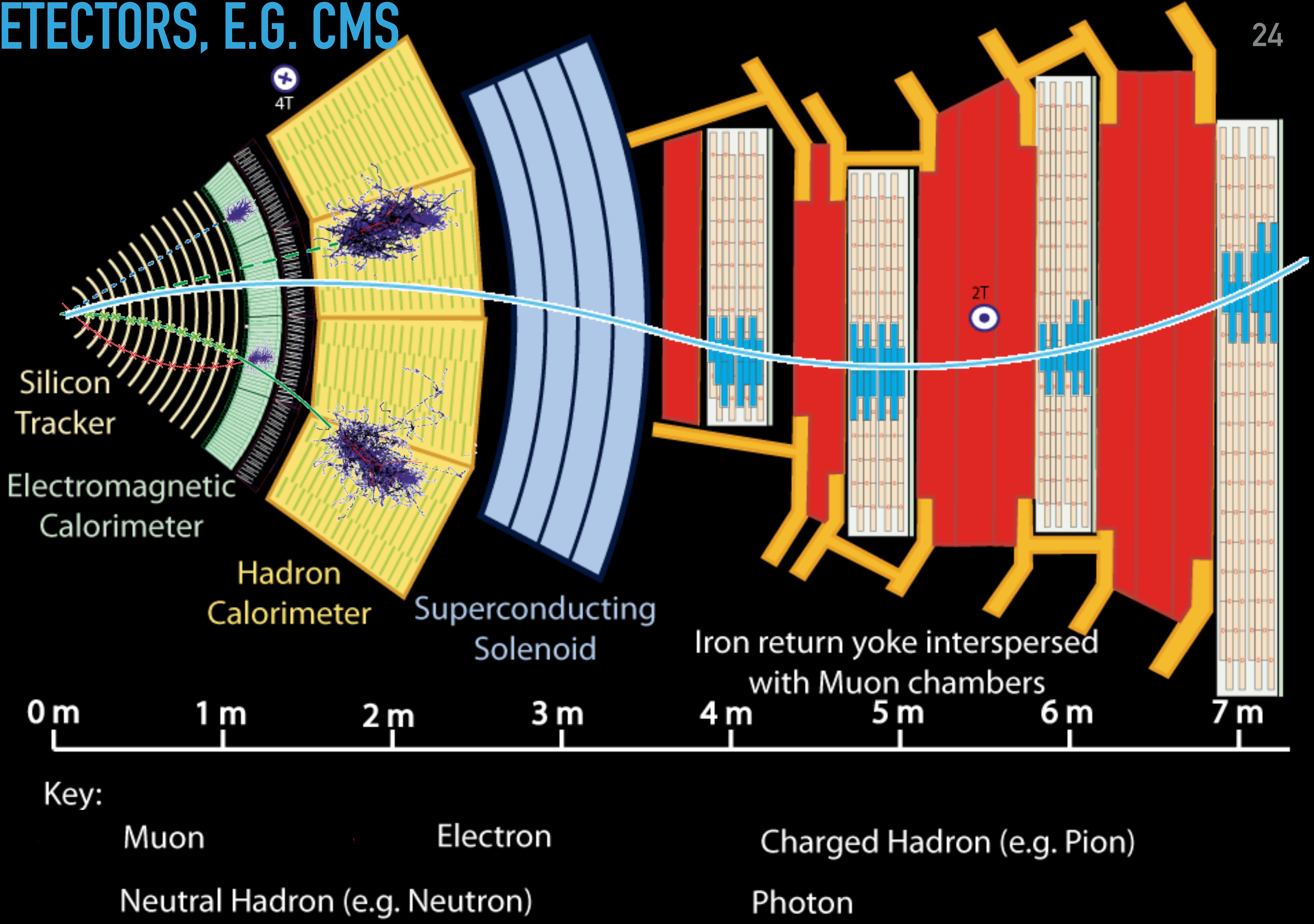
OUTLOOK



MULTILAYERED DETECTORS, E.G. CMS



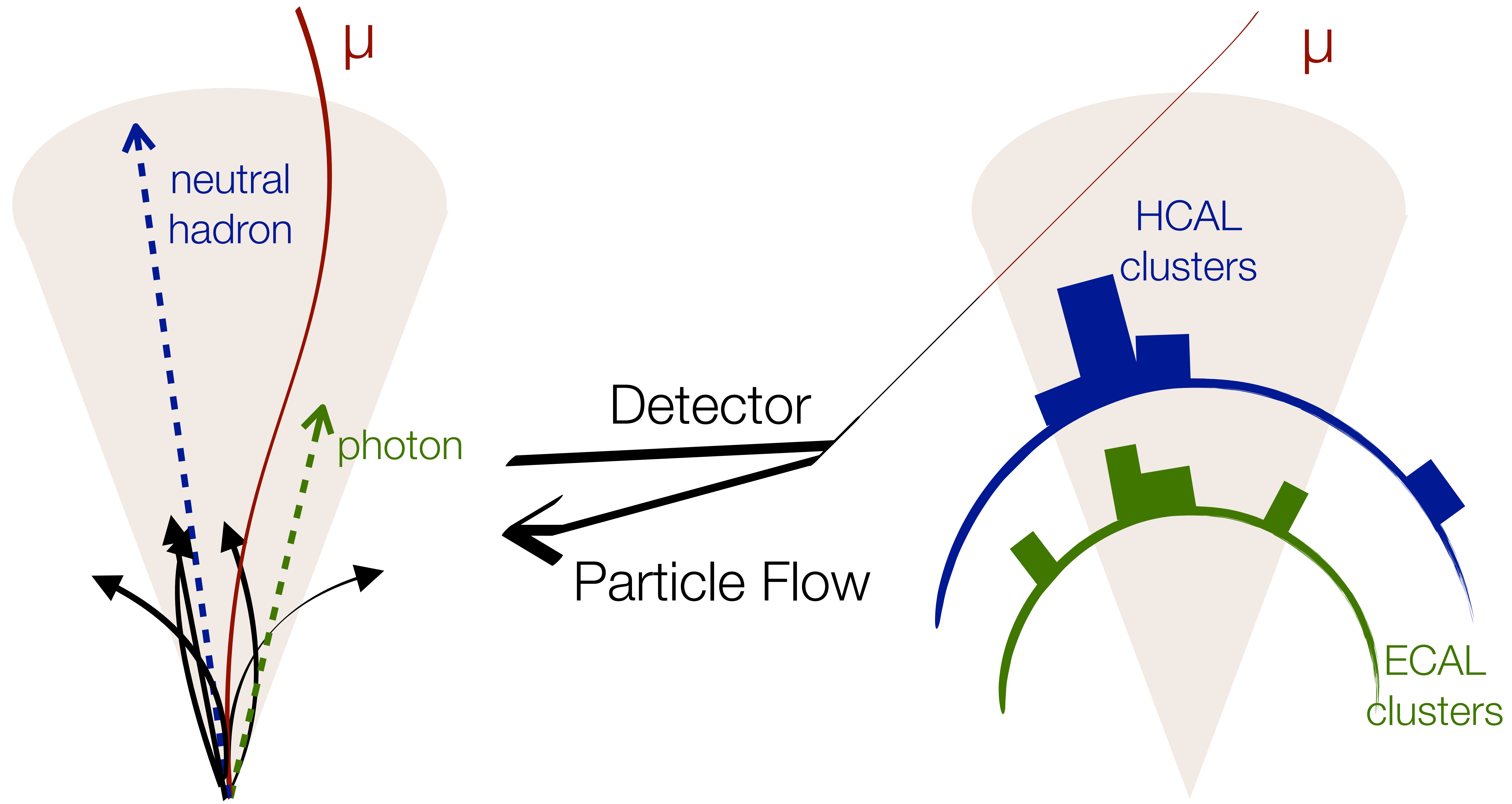
Current and future multilayered detectors…



Require complex pattern recognition

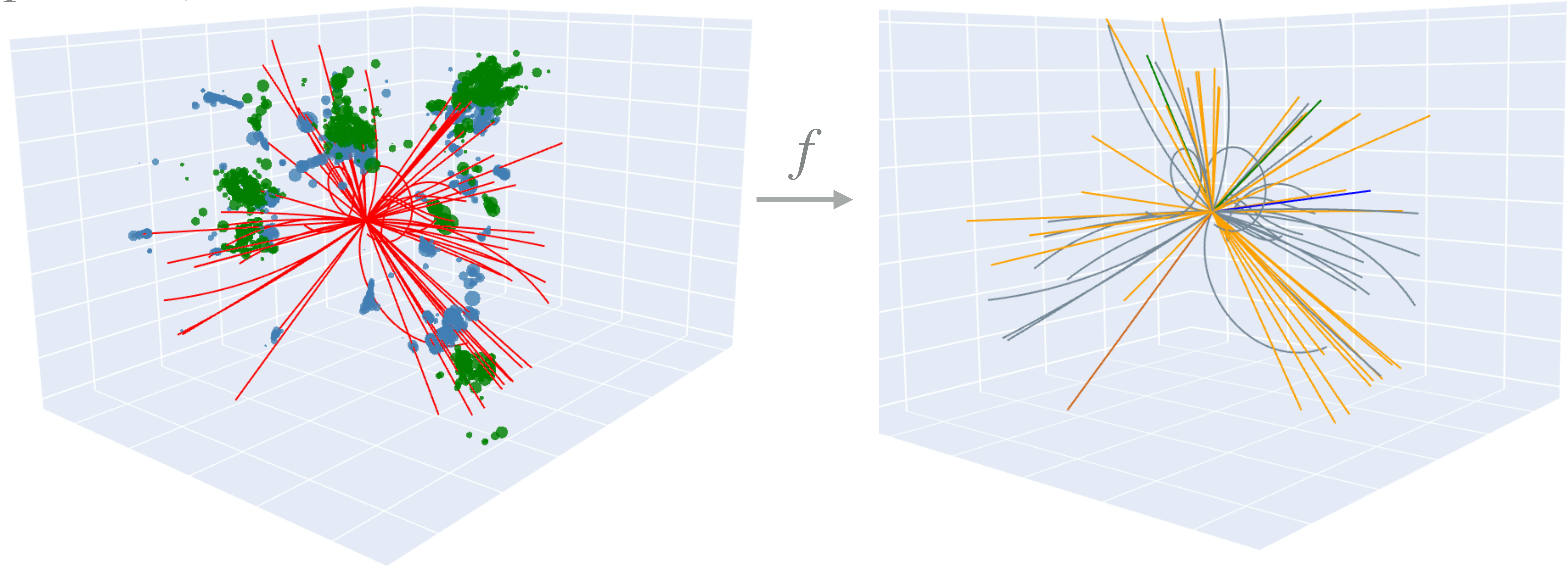
PARTICLE-FLOW RECONSTRUCTION

- Particles interact with detector, leaving energy deposits and tracks
- Combination of multimodal information from complementary detectors to produce particle-level interpretation of the event based on complex, hand-tuned heuristics



PARTICLE-FLOW AS A MACHINE-LEARNING TASK

- Can we instead formulate PF as an ML task (naturally "tunable" through re-training and portable to new hardware)?
- Learn a "set-to-set" function $f: X \to Y$, where $\{\text{tracks, energy clusters}\} \in X$ and $\{\text{particles}\} \in Y$



$\{\text{tracks, energy clusters}\} \in X$

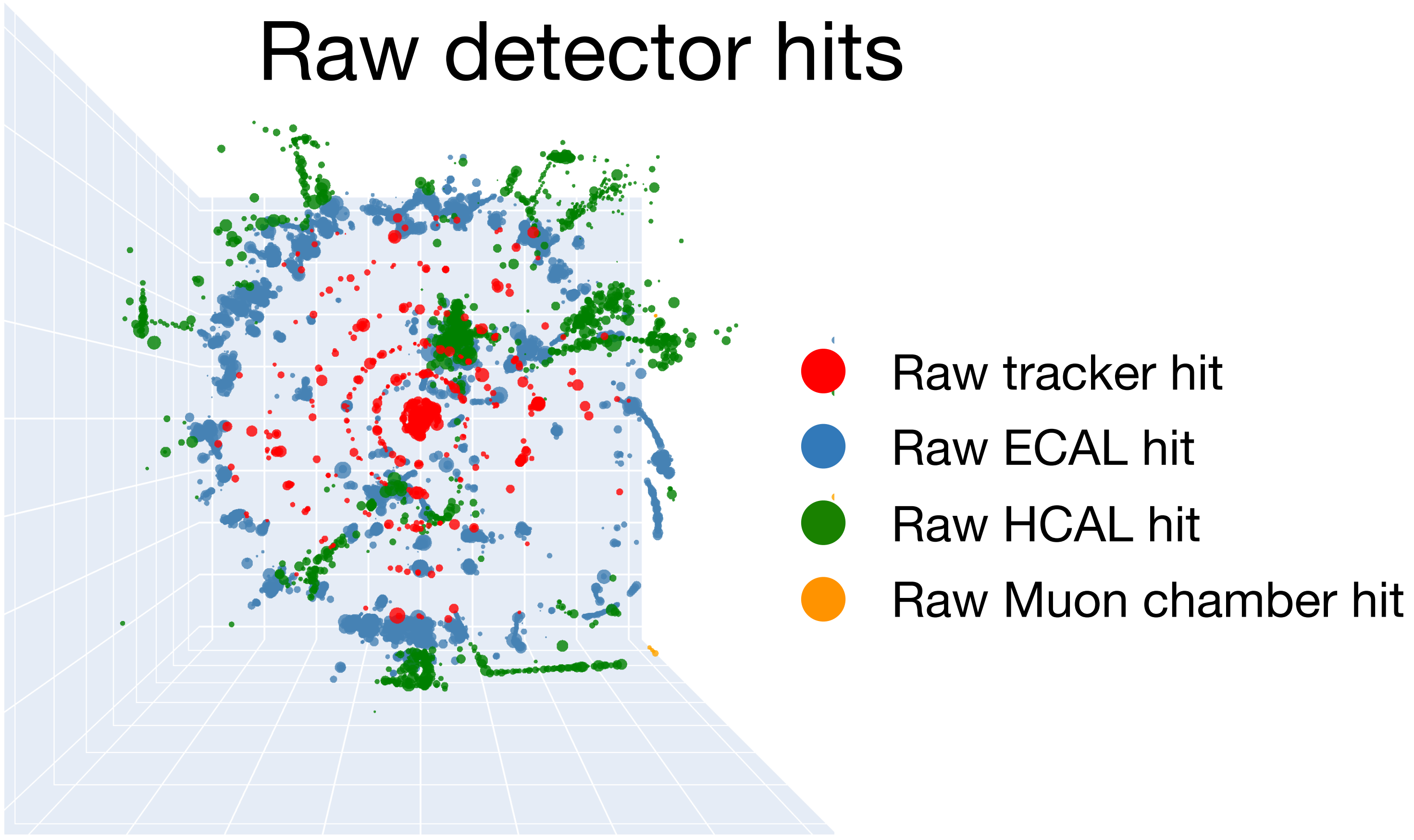
$\{\text{particles}\} \in Y$

OPEN DATASET FOR ML RECONSTRUCTION STUDIES

doi:10.5281/zenodo.8260741
https://www.coe-raise.eu/od-pfr

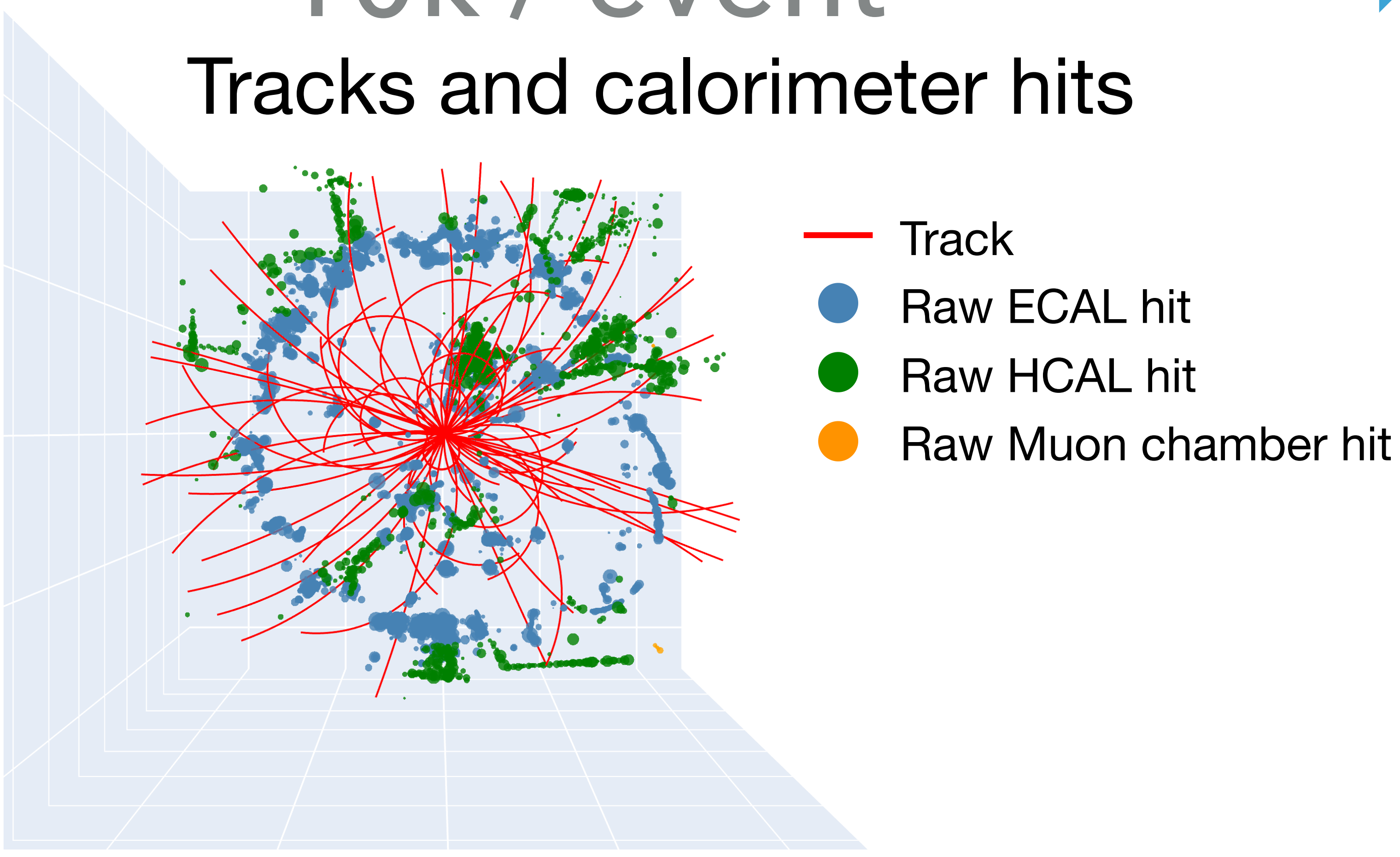
- 2.5 TB, 6 million events total

~100k / event

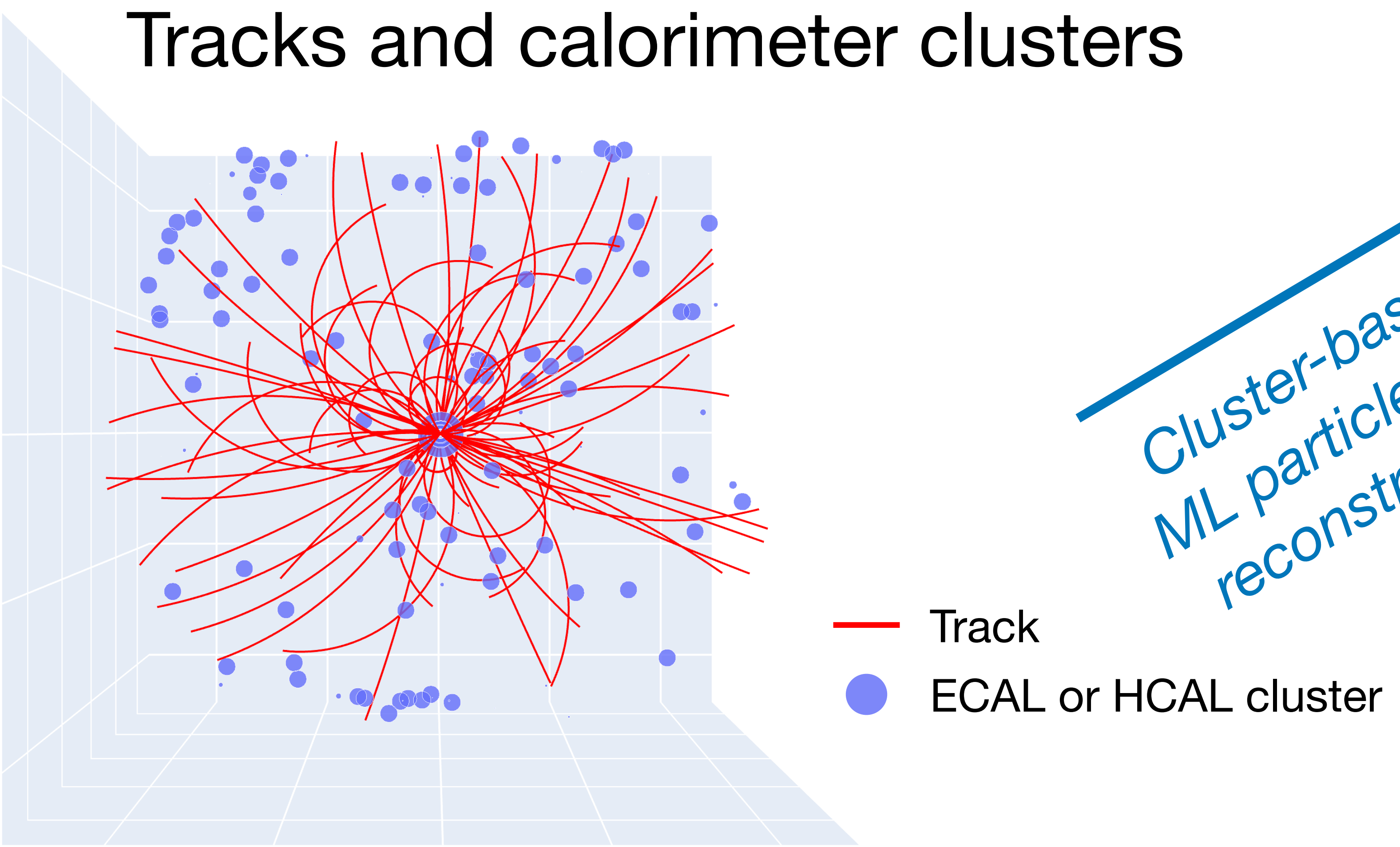


Charged particle tracking

~10k / event

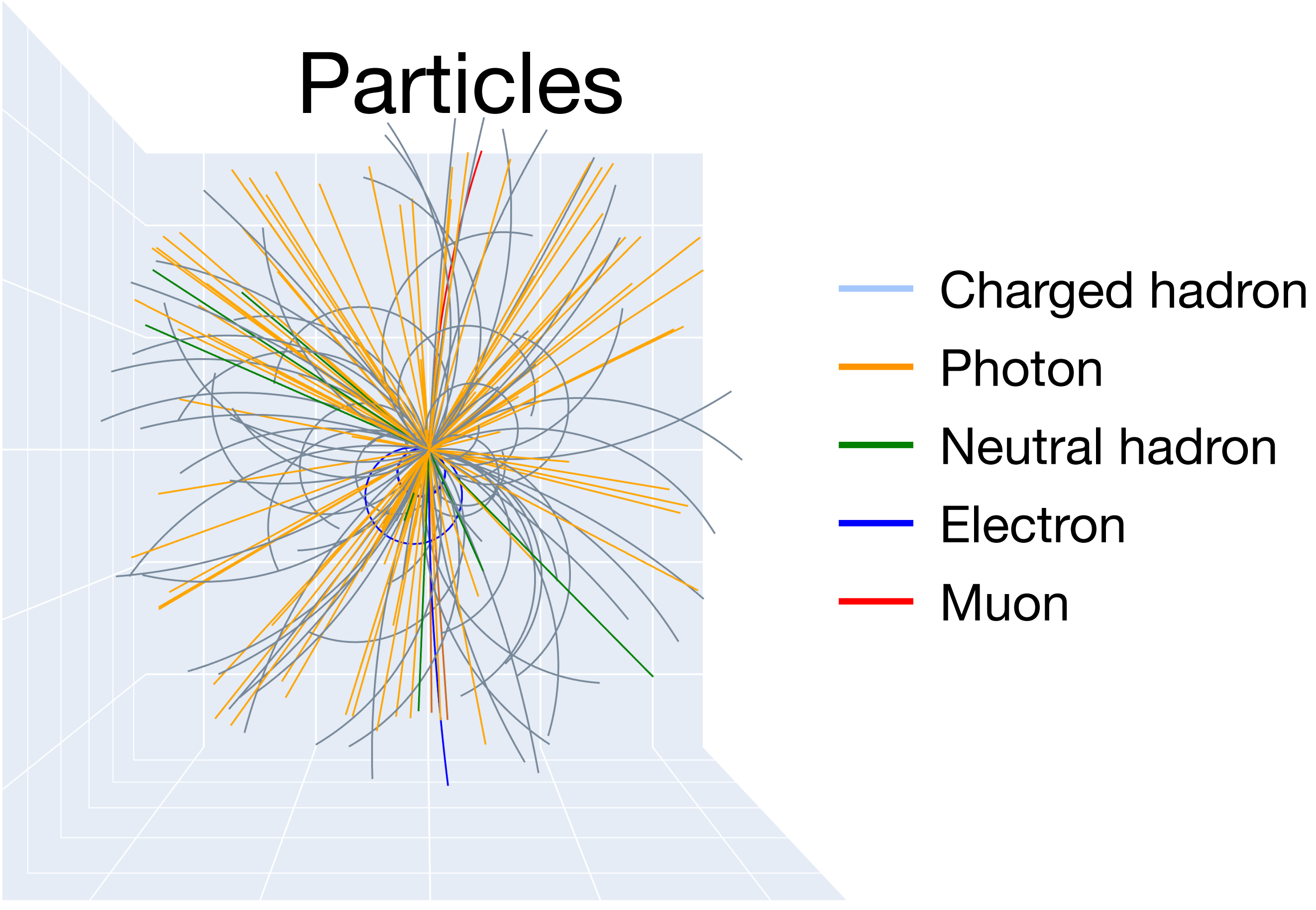


Calorimeter clustering



~300-500 / event

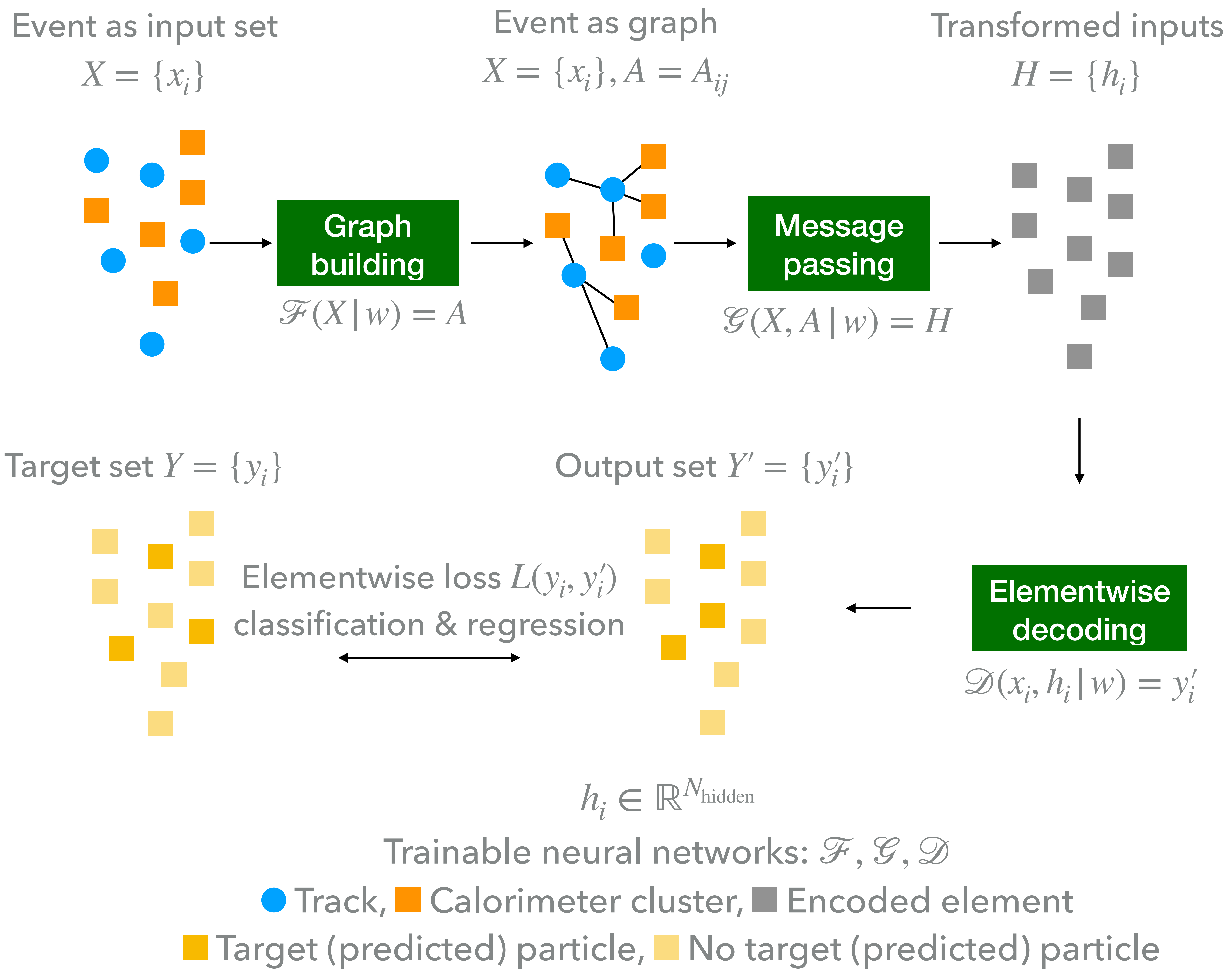
Cluster-based ML particle-flow reconstruction



~100-300 / event

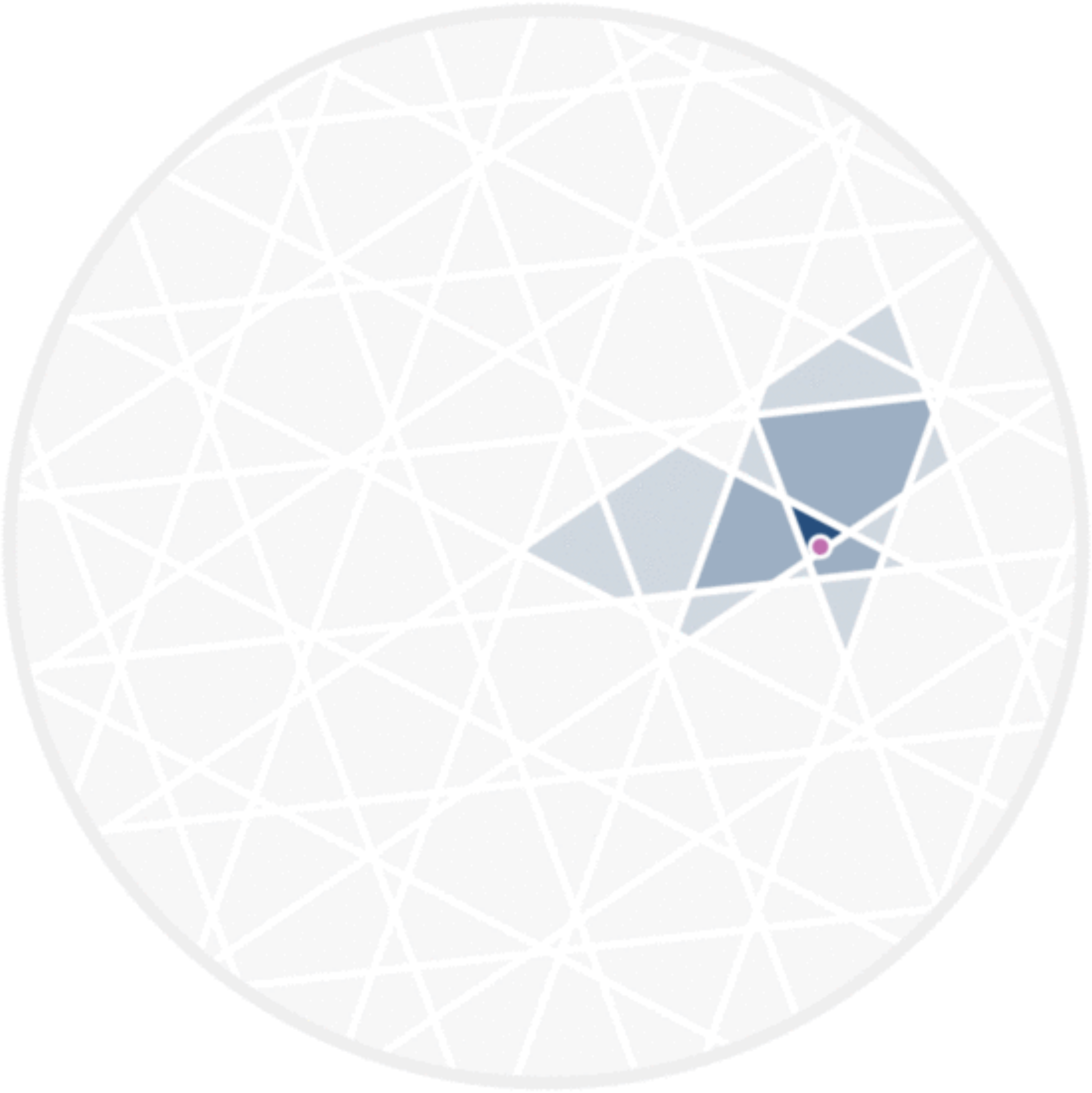
GRAPH NEURAL NETWORK APPROACH

- Convert input set to a locally, **sparsely connected** graph
- Message-passing NN to transform features
- Decode transformed inputs elementwise
- (During training) Compare to target set, optimize weights



T. Neylon, https://unboxresearch.com/articles/lsh_post1.html

Locality-sensitive hashing reduces graph-building complexity



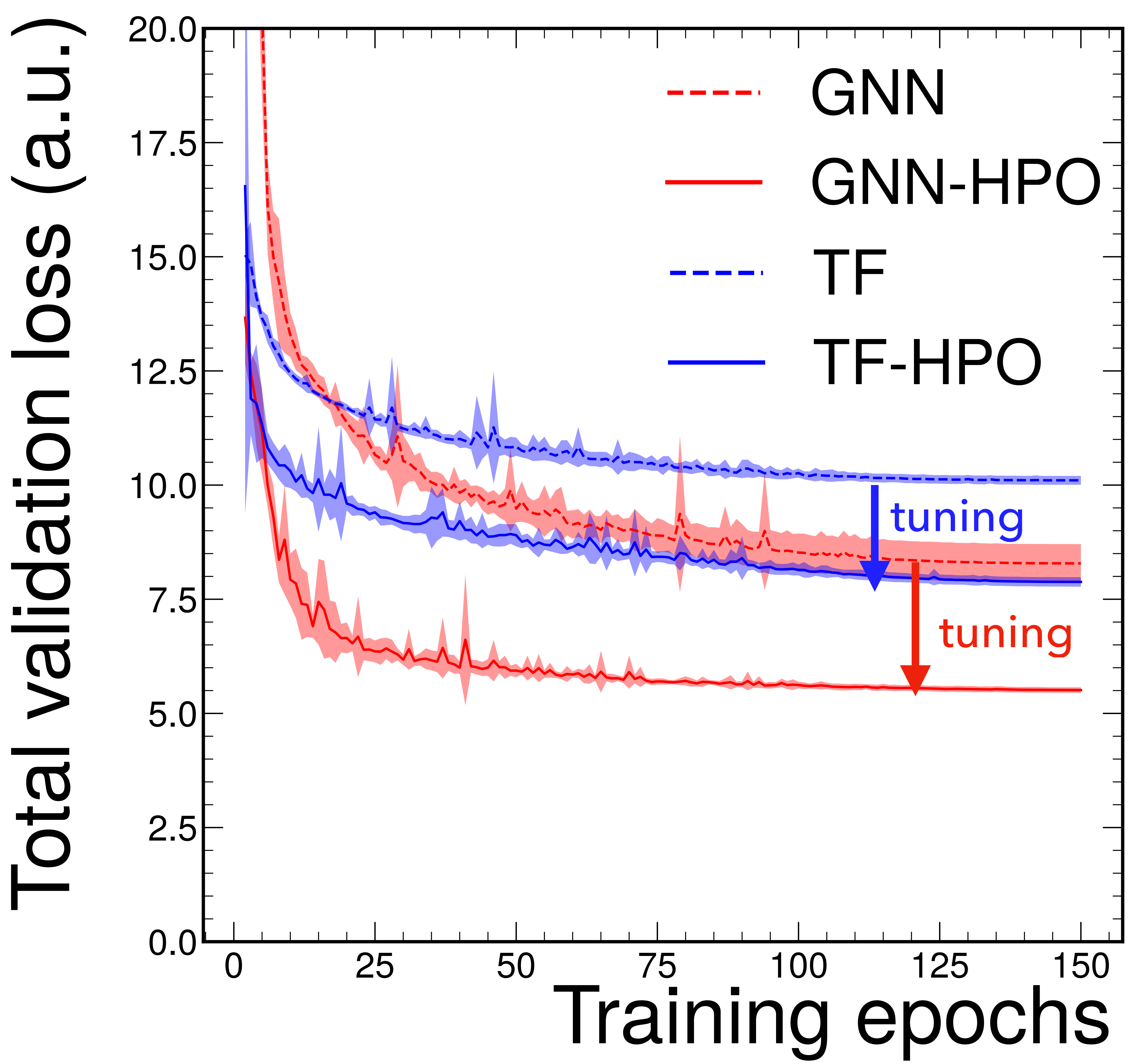
5 random hash functions, where darkest blue = 5 hash collisions, lightest blue = 3 hash collisions

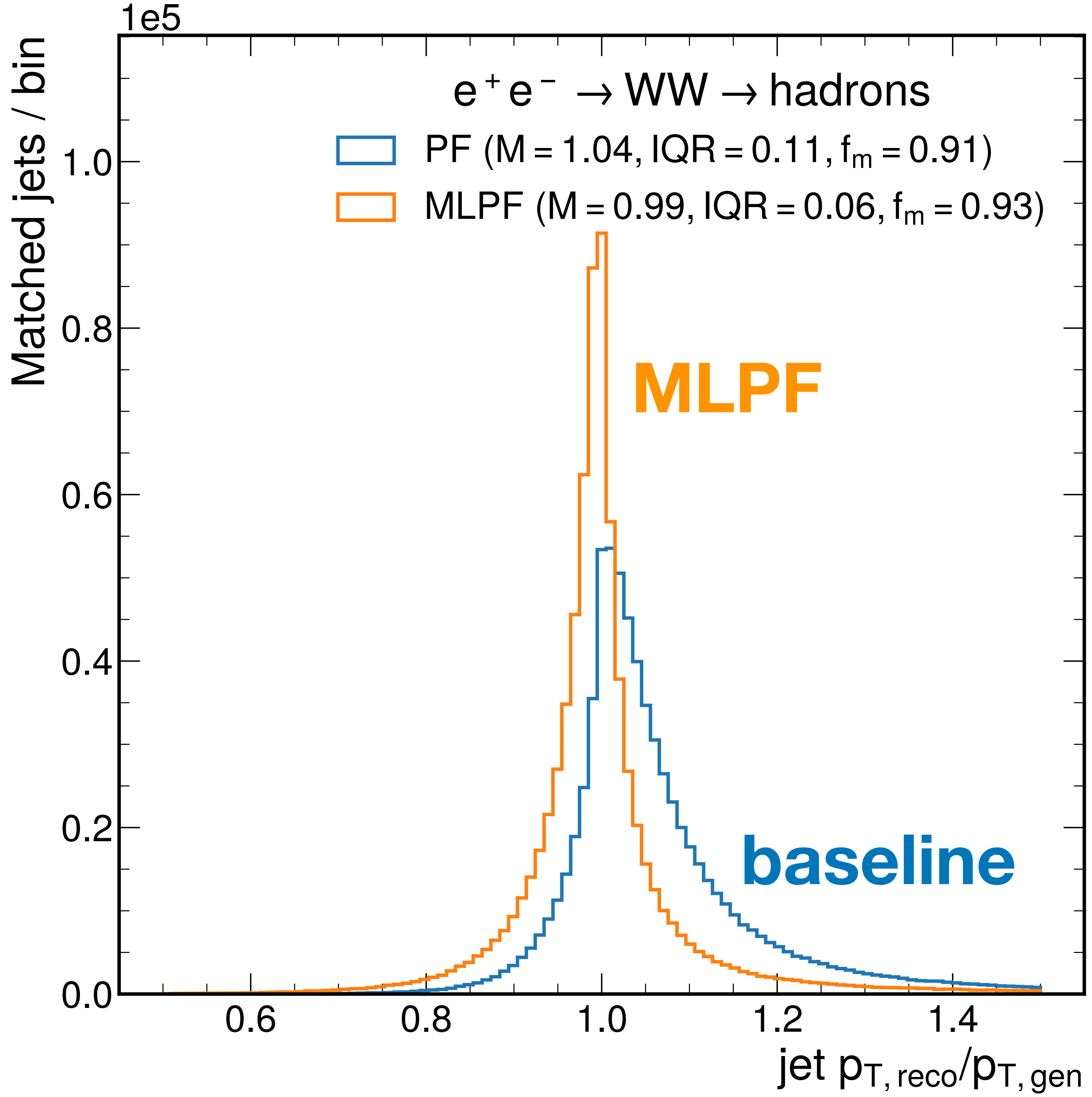
HYPERPARAMETER OPTIMIZATION ON HPC

- Hyperparameter optimization requires large compute

IMPACT OF TUNING

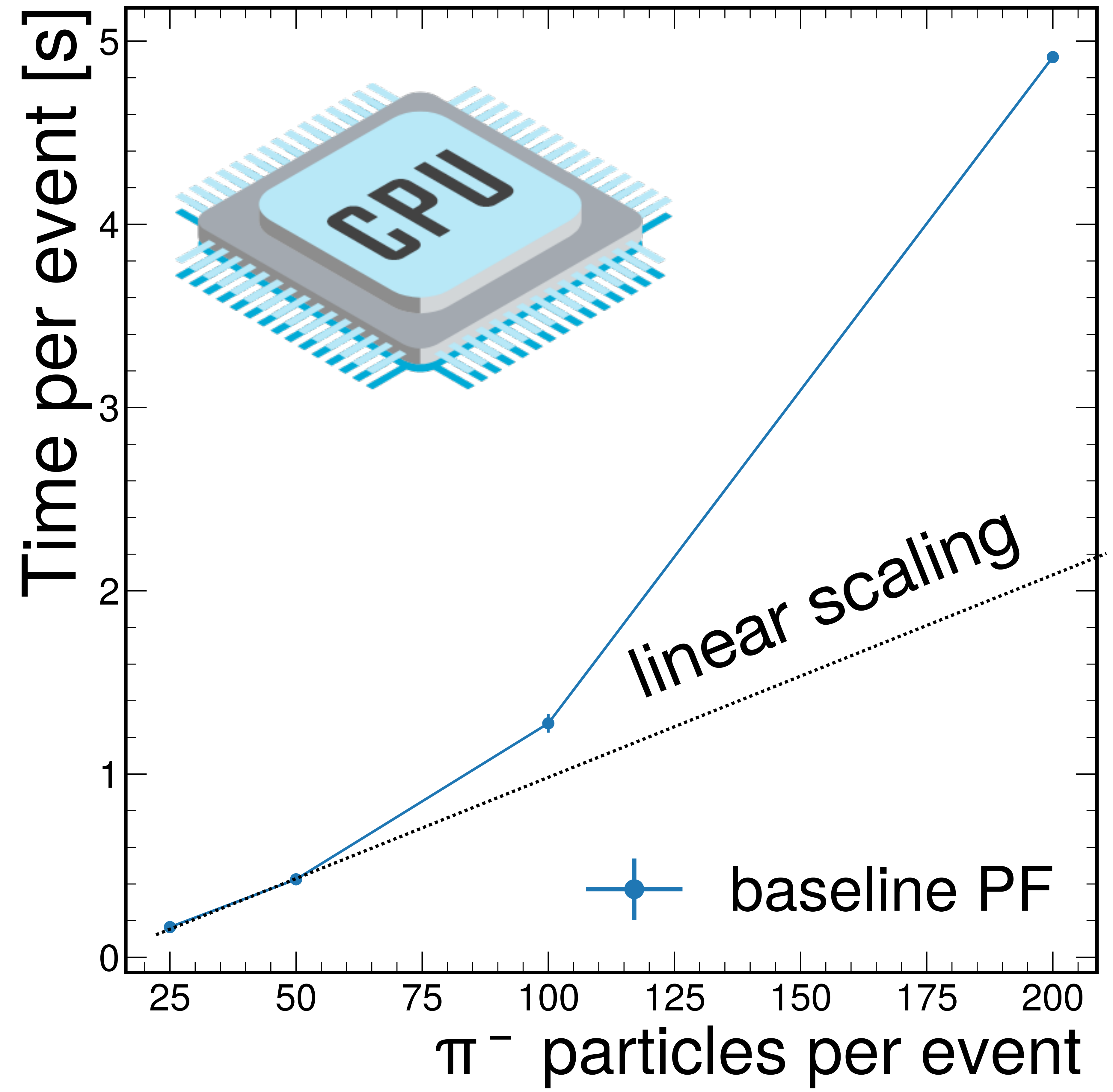
- Tuning improves particle-level performance dramatically
- Though we optimize a particle-level loss, also achieve better energy resolution

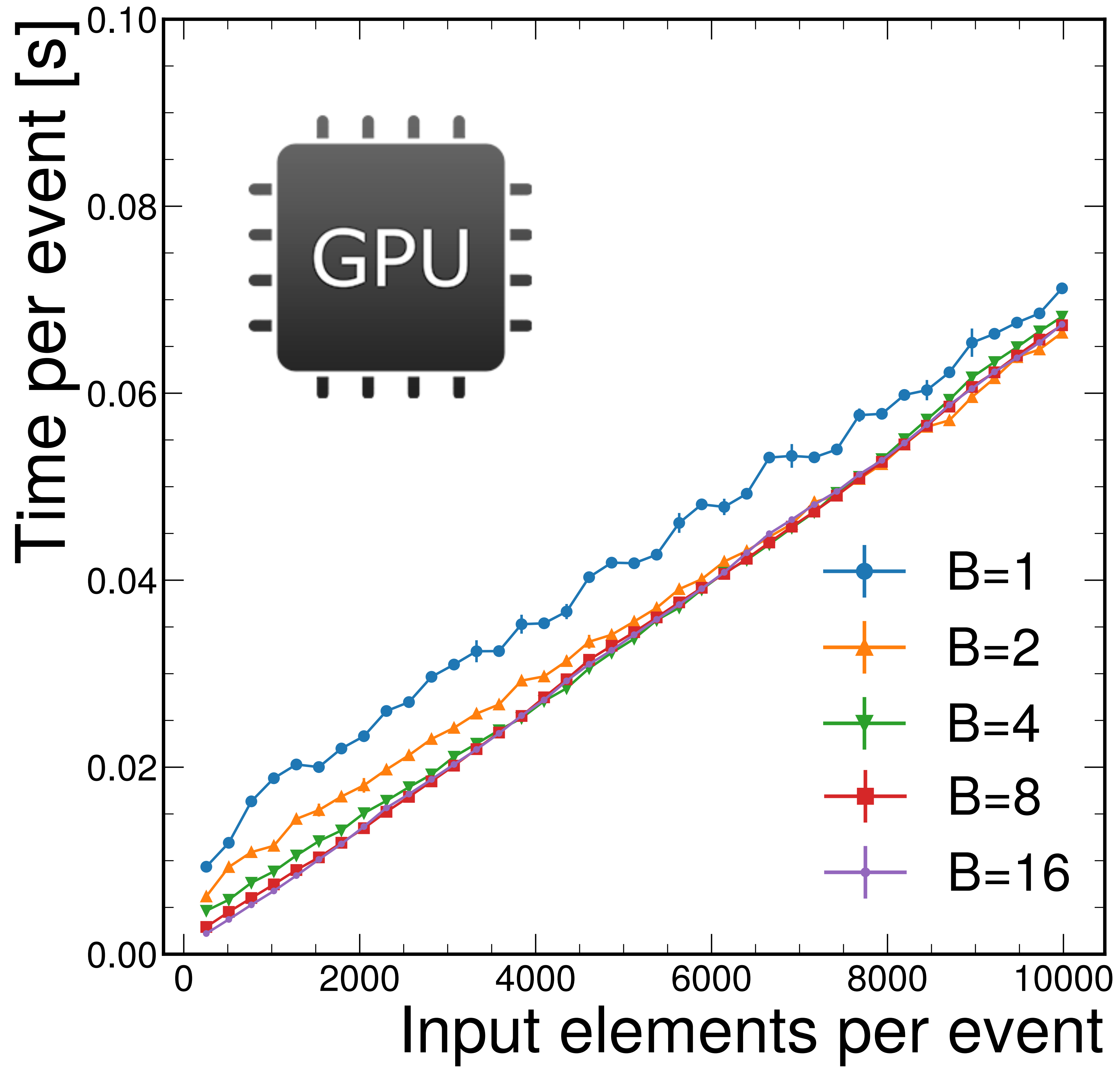




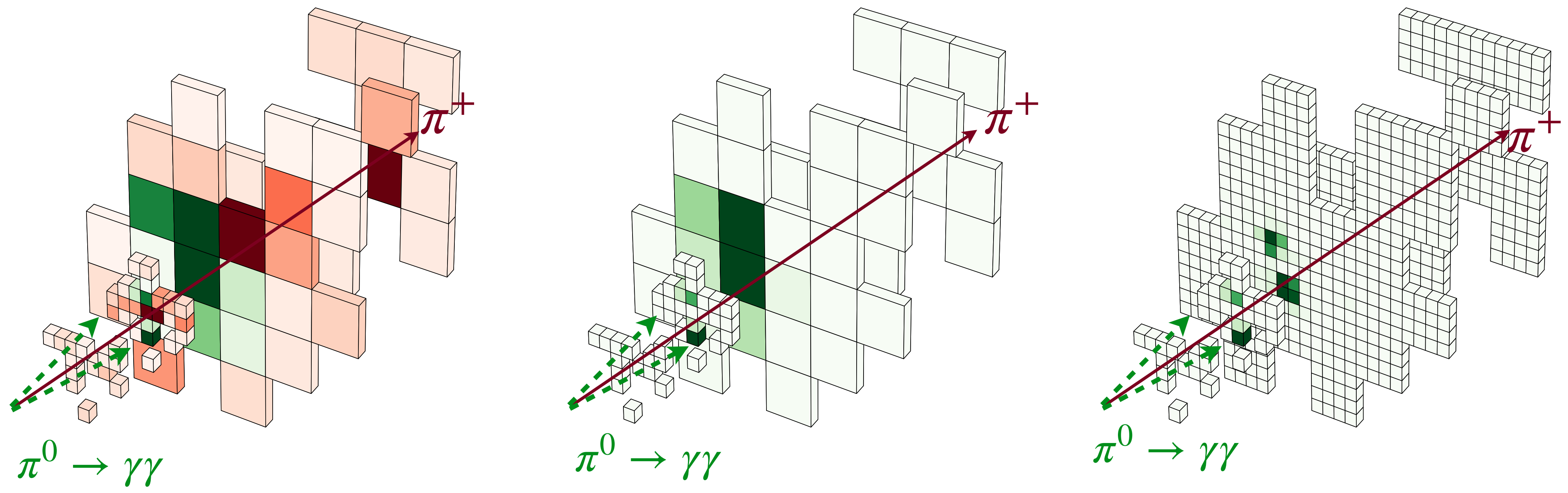
SCALING

- Baseline algorithm runs only on CPU, scales ~quadratically, runs in seconds
- ML model scales linearly, runs in milliseconds on a consumer 8 GB GPU





VARIETY OF APPROACHES

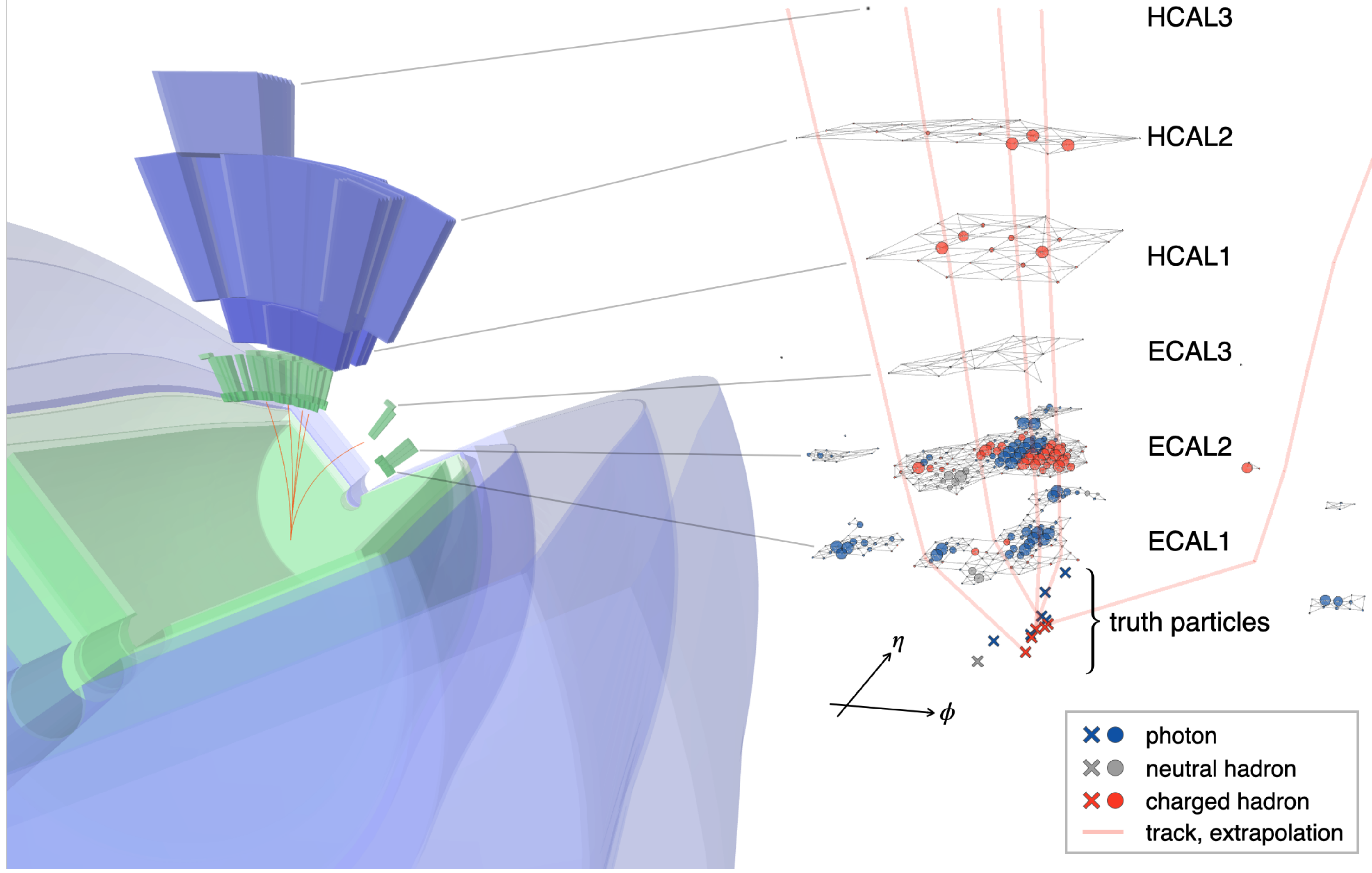


Towards a Computer Vision Particle Flow ⋆

Francesco Armando Di Bello[a,3], **Sanmay Ganguly**[b,1], **Eilam Gross**[1], **Marumi Kado**[3,4], **Michael Pitt**[2], **Lorenzo Santi** [3], **Jonathan Shlomi**[1]

[1]Weizmann Institute of Science, Rehovot 76100, Israel
[2]CERN, CH 1211, Geneva 23, Switzerland
[3]Università di Roma Sapienza, Piazza Aldo Moro, 2, 00185 Roma, Italy e INFN, Italy
[4]Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405, Orsay, France

Eur. Phys. J. C 81, 107 (2021)



Reconstructing particles in jets using set transformer and hypergraph prediction networks

Francesco Armando Di Bello [1,a], **Etienne Dreyer** [2,b], **Sanmay Ganguly** [3], **Eilam Gross** [2], **Lukas Heinrich** [4], **Anna Ivina** [2], **Marumi Kado** [5,6], **Nilotpal Kakati** [2,c], **Lorenzo Santi** [6], **Jonathan Shlomi** [2], **Matteo Tusoni** [6]

[1] INFN and University of Genova
[2]Weizmann Institute of Science
[3]ICEPP, University of Tokyo
[4]Technical University of Munich
[5]Max Planck Institute for Physics
[6]INFN and Sapienza University of Rome

Eur. Phys. J. C 83, 596 (2023)

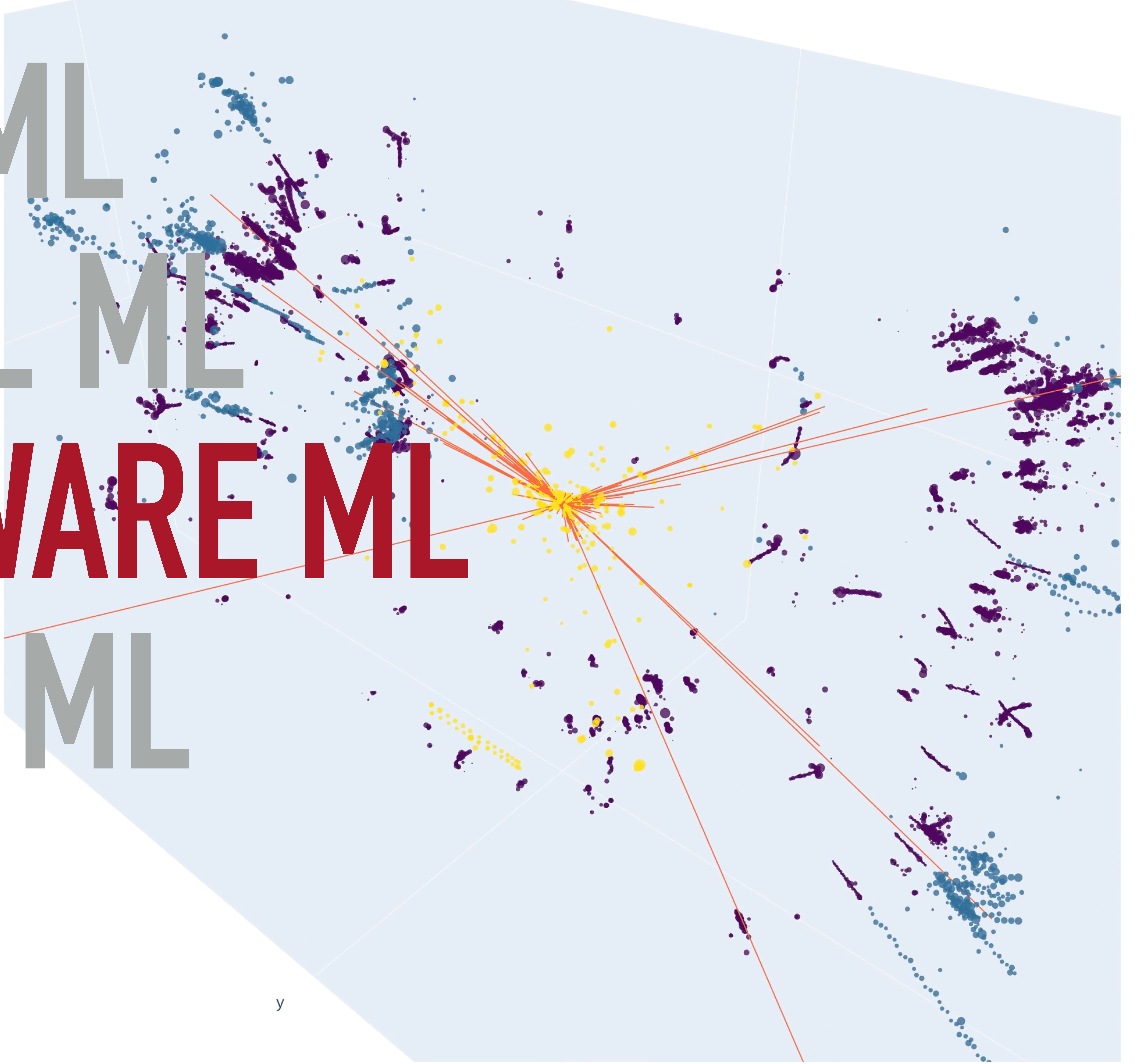
ULTRAFAST ML

MULTIMODAL ML

PHYSICS-AWARE ML

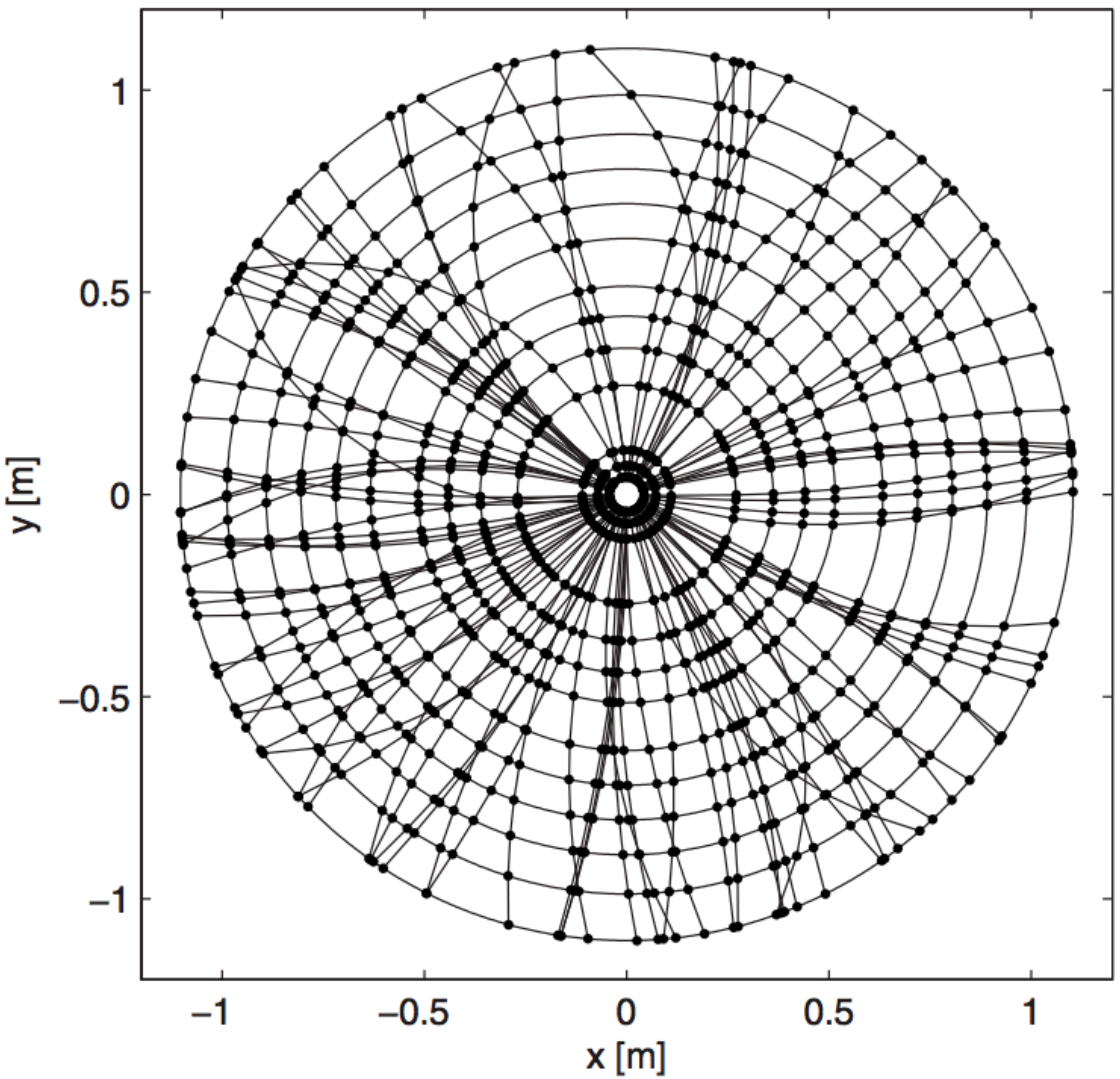
GENERATIVE ML

OUTLOOK



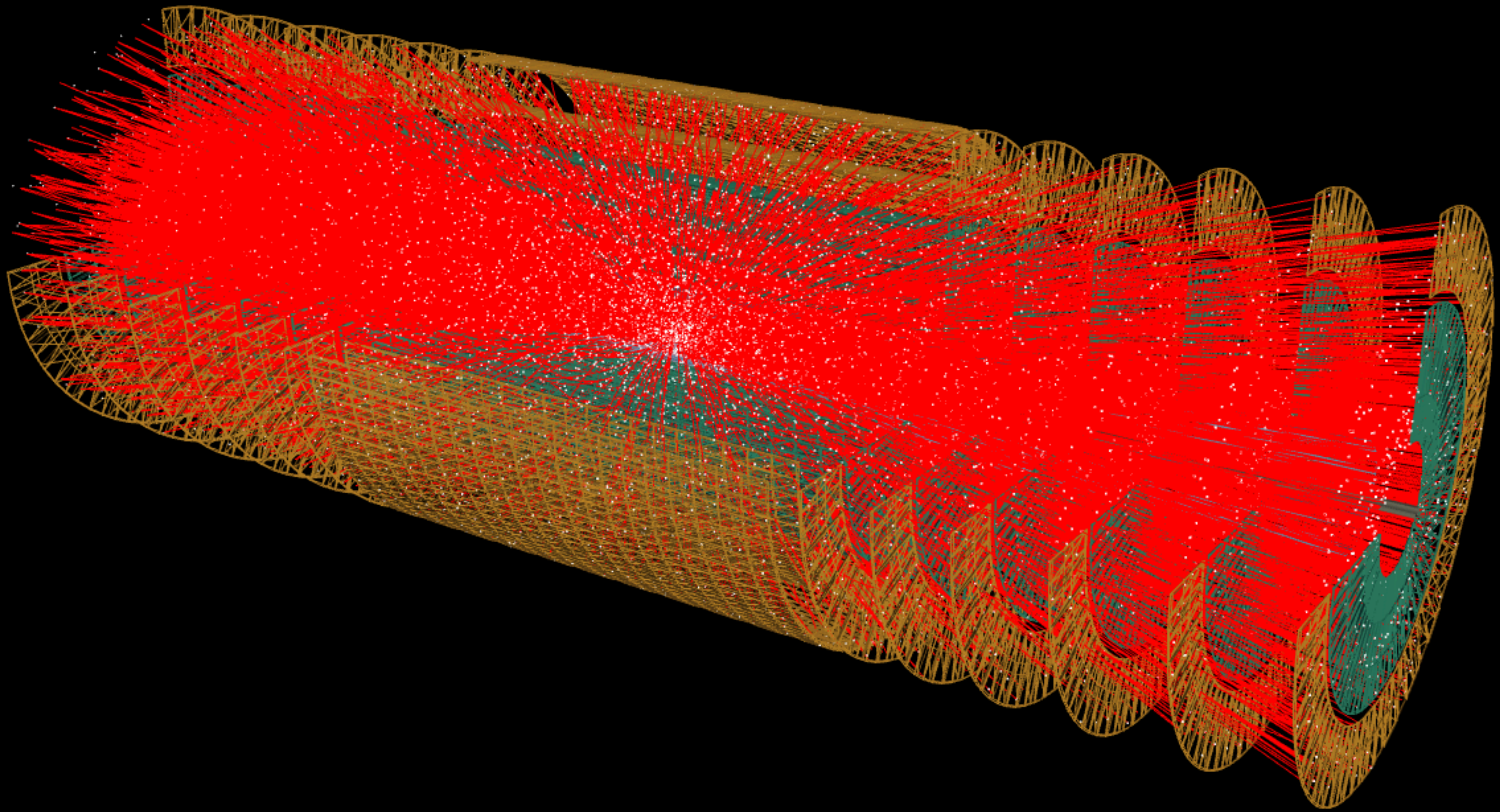
TRACKING (CONNECTING THE DOTS)

- Particle tracking is a classic pattern recognition task
- From a set of hits sampled sparsely in 3D, reconstruct the helical trajectories of particles
- Traditional algorithms scale worse than N^2 in the number of hits N
- How about GNNs and transformers?



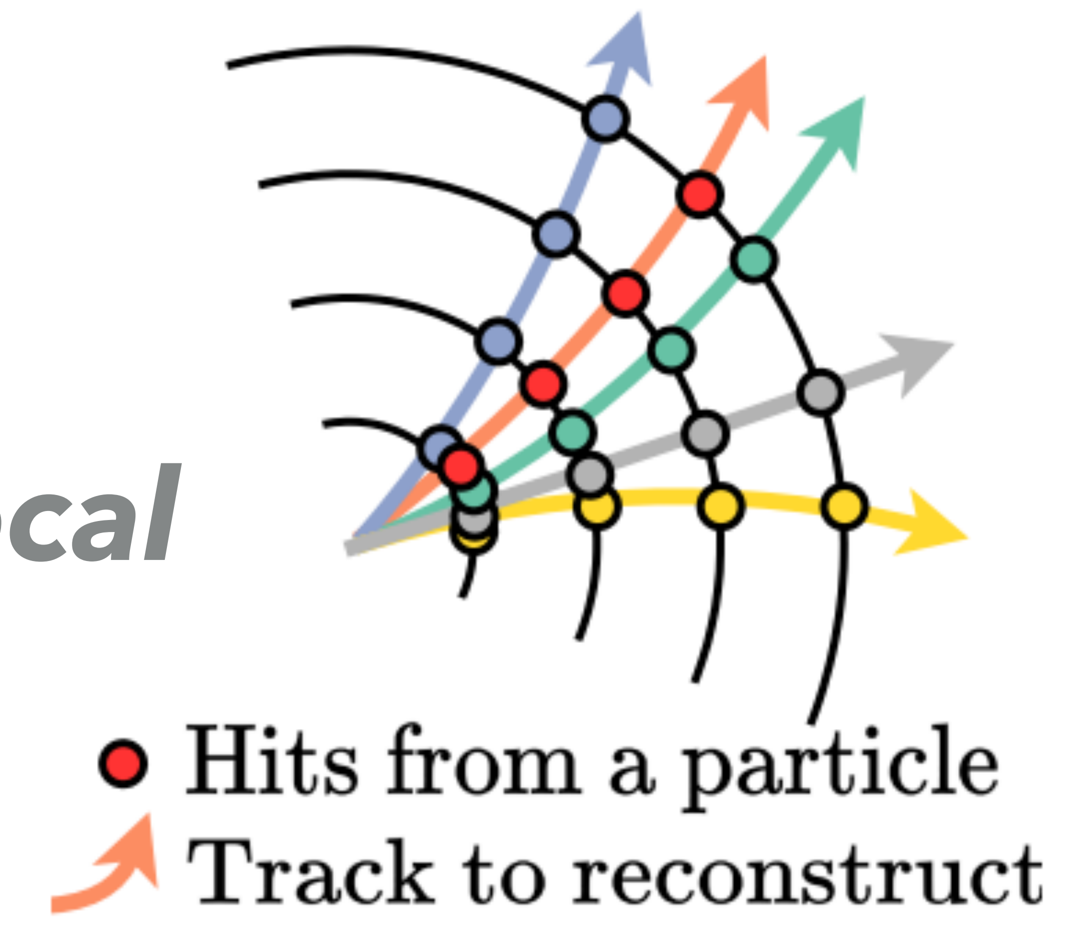
TRACKML DATASET

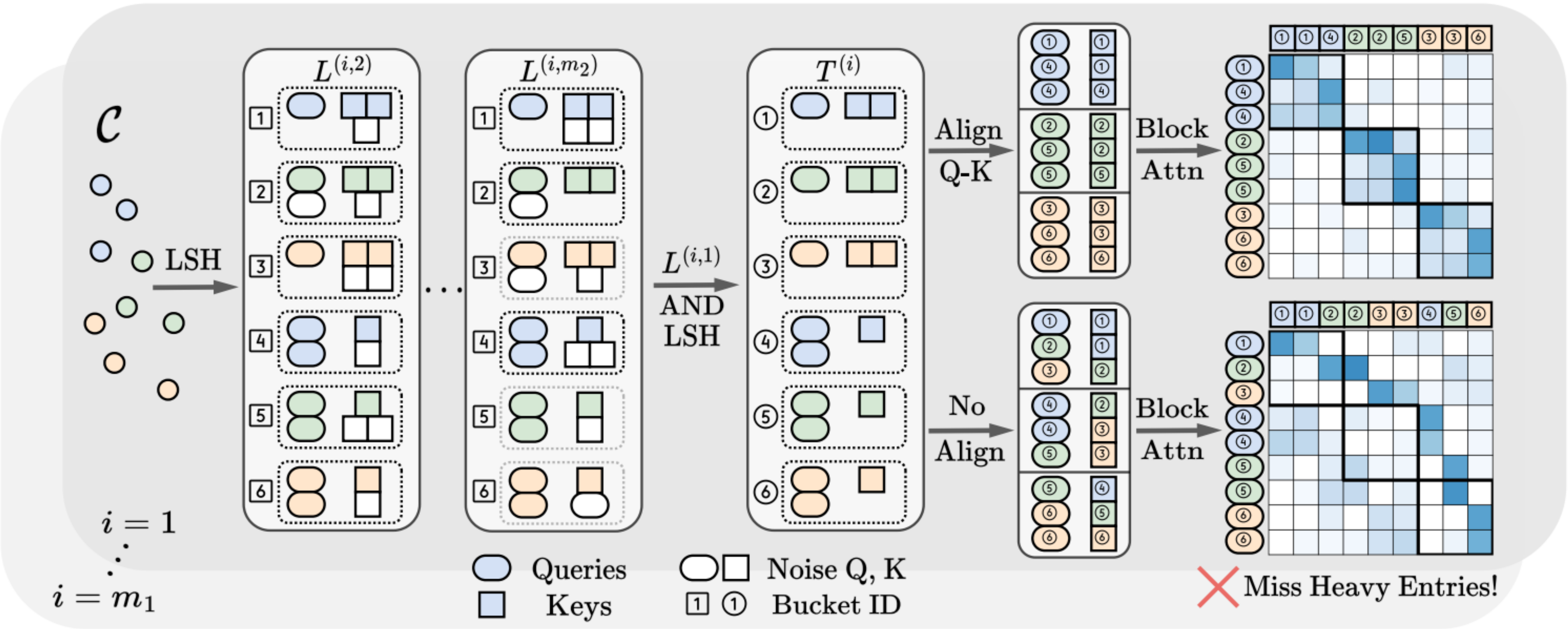
- More realistic public TrackML dataset used for 2018 Kaggle competition has O(100k) hits and O(10k) tracks per event
  - $O(N^2)$ attention or graph building is too slow!

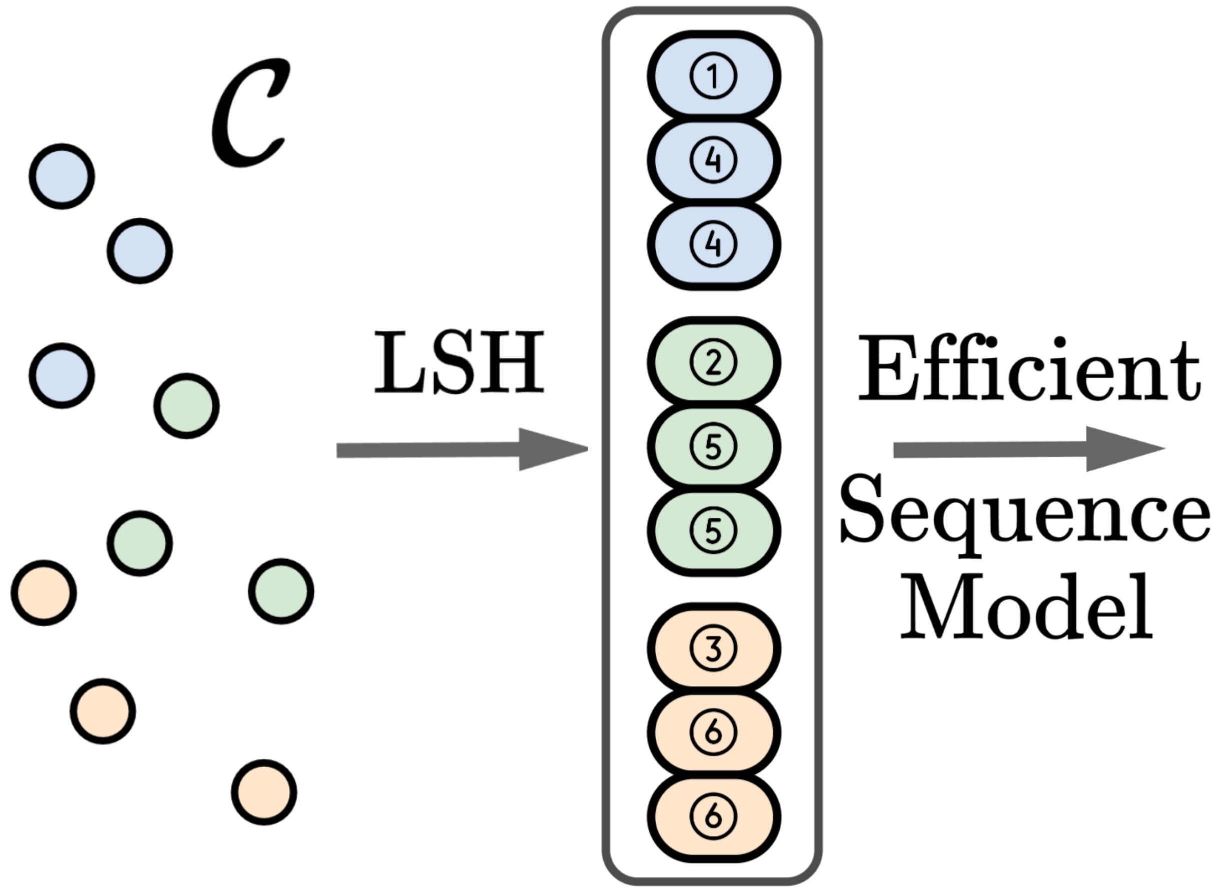


LSH-BASED EFFICIENT POINT TRANSFORMER (HEPT)

- HEPT: an efficient point transformer based on OR & AND LSH
  - No graph construction; only regular computations
  - Assign hash codes using OR & AND E^{2}LSH; **physics-aware *local inductive bias:*** nearby hits in detector share 1D hash codes
  - Sort items based on hash codes; block-diagonal attention



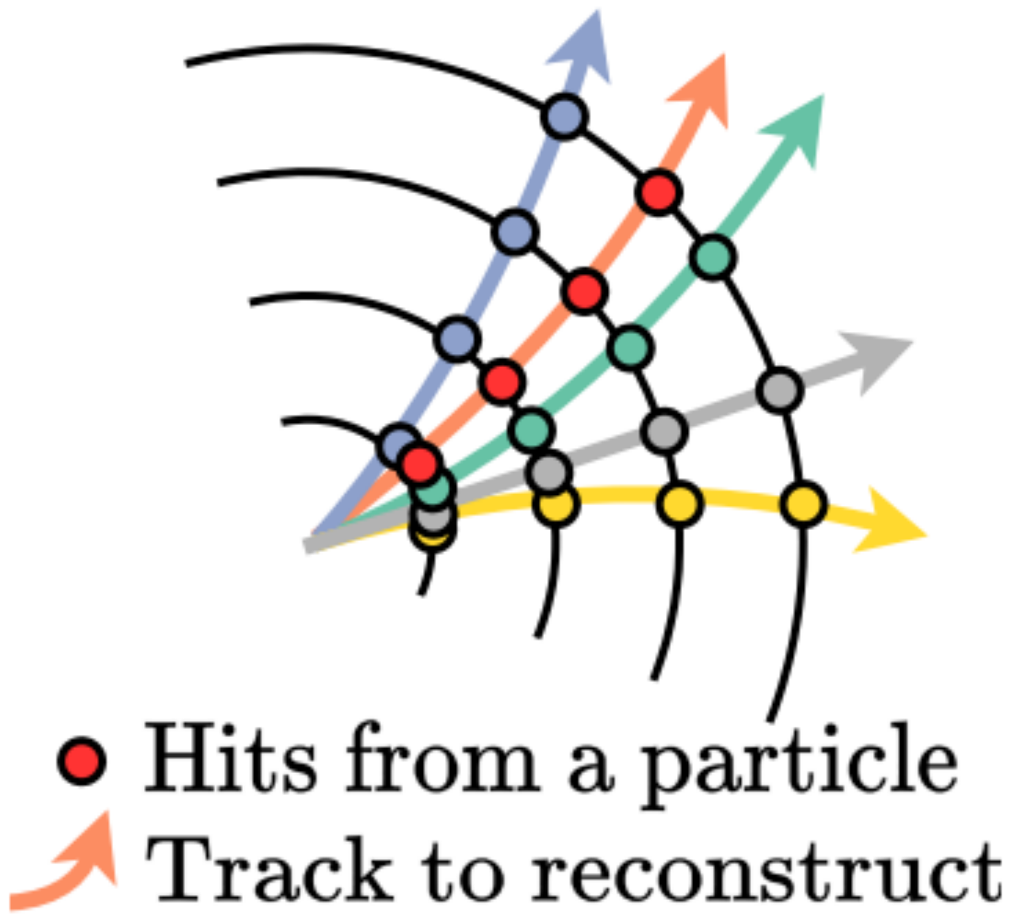




Key idea: HEPT projects point clouds to 1D sequences using ***physics-aware local inductive bias***

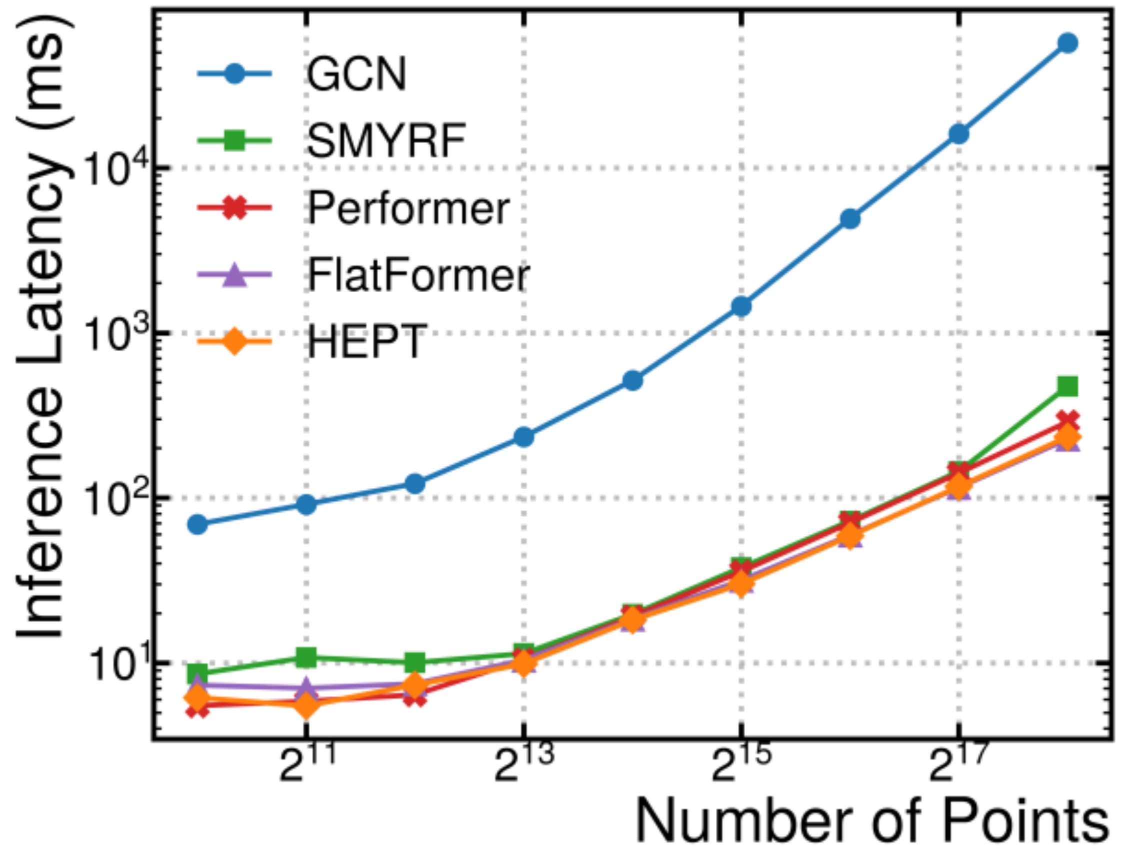
EMPIRICAL RESULTS

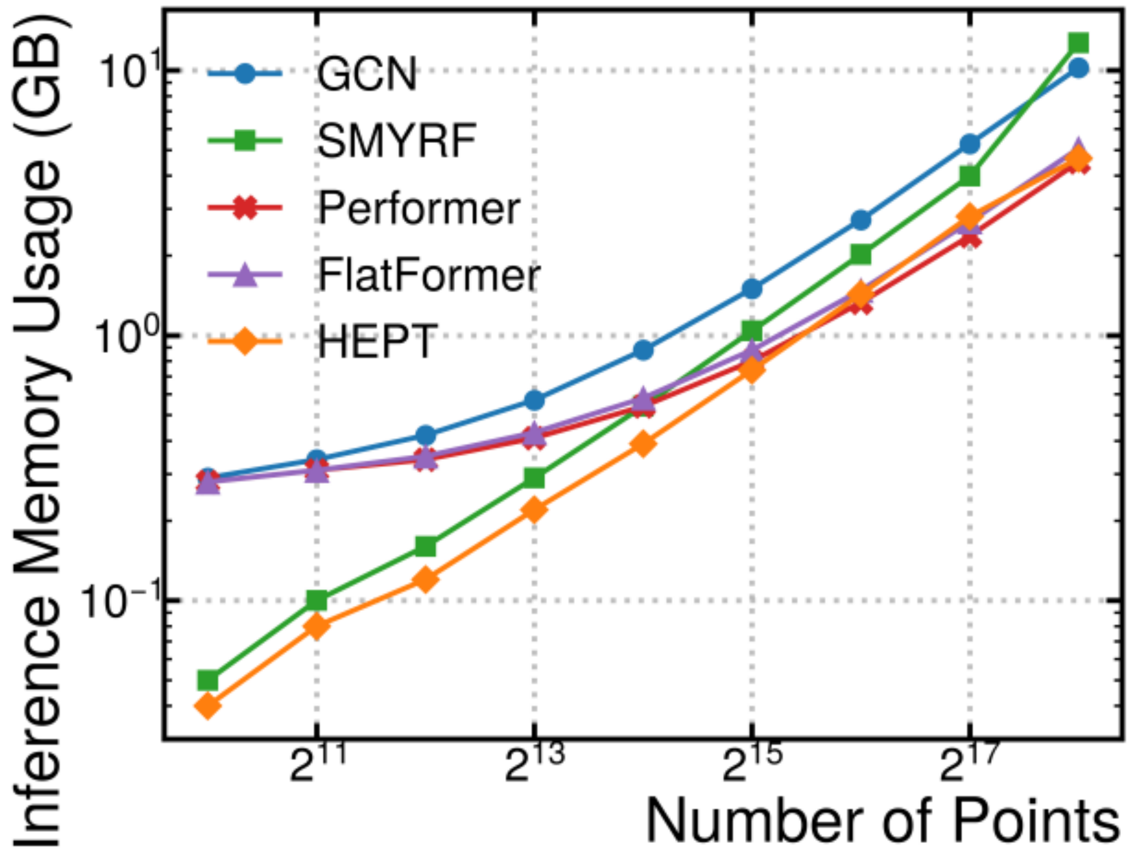
- Tracking as a representation learning task: learn close embeddings for hits originating from the same particle



	Tracking-6k (AP@k)	Tracking-60k (AP@k)
Random	5.88	5.71
SOTA GNNs	**91.00**[‡]	**90.89**[‡]
Reformer	72.37	72.47
SMYRF	72.98	71.18
HyperAttn	71.49	70.22
Performer	73.17	72.07
FLT	72.55	71.45
ScatterBrain	73.35	72.06
PointTrans	72.33	70.81
FlatFormer	74.22	70.23
GCN	79.61	75.38
DGCNN	**90.74**	**88.66**
GravNet	90.11	87.99
GatedGNN	80.98	78.42
Performer-k_{HEPT}	71.97	69.20
SMYRF-k_{HEPT}	83.19	71.04
FlatFormer-k_{HEPT}	88.18	85.06
HEPT	**92.66**[†]	**91.93**[†]

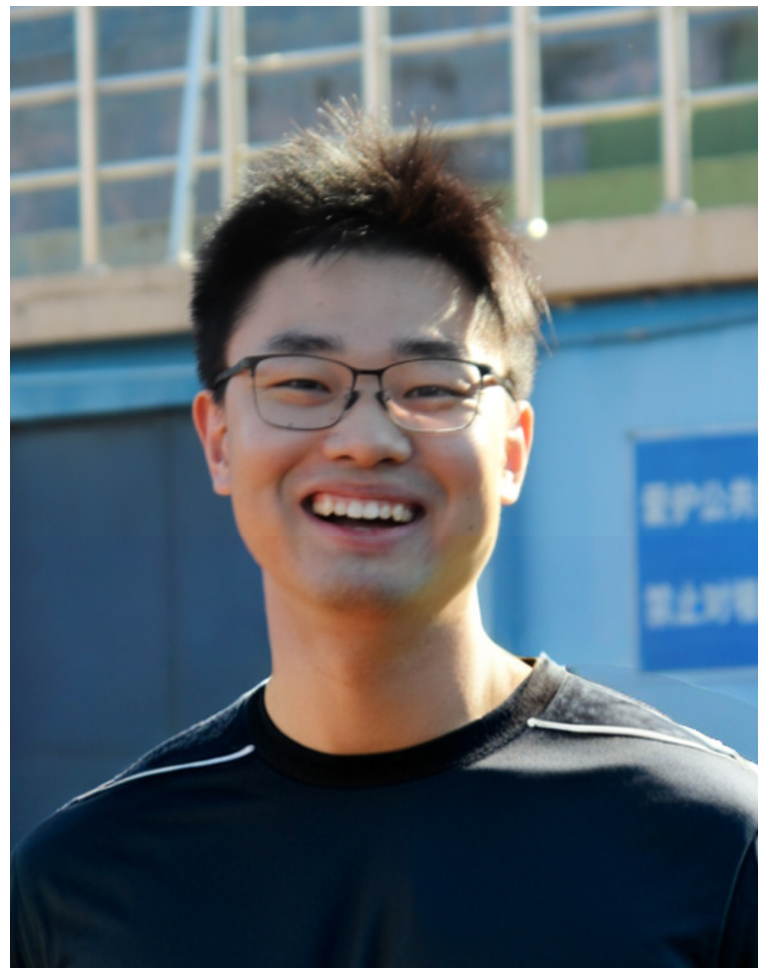
- HEPT achieves SOTA performance and achieves over ***100x speedup*** on GPUs compared to GNNs on Tracking-60k (60k hits/event)

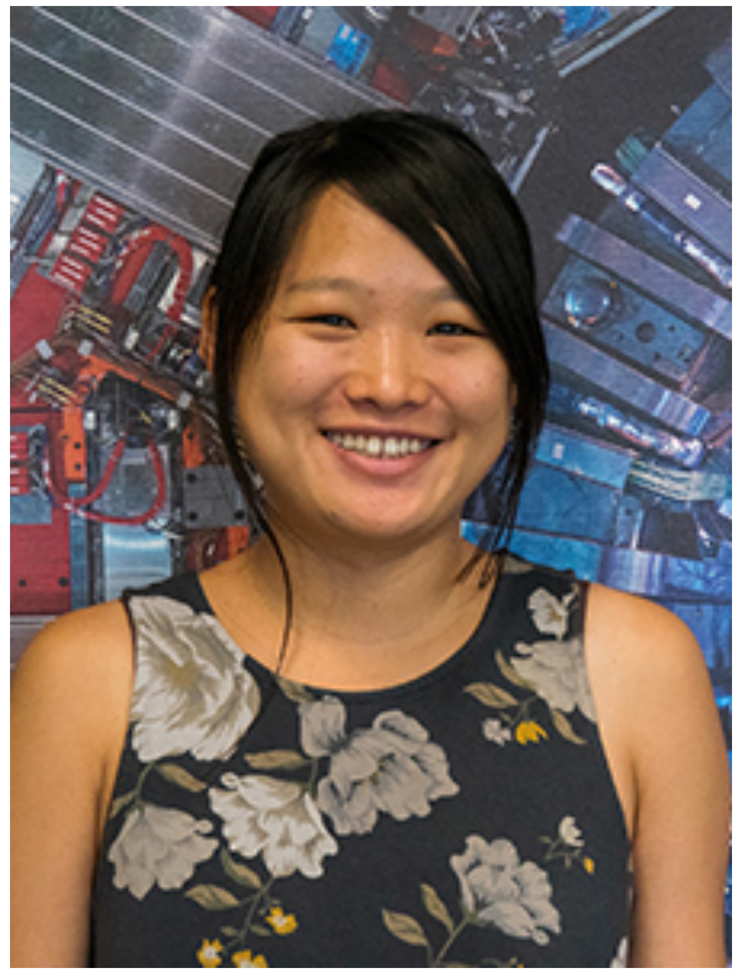




HEAR MORE TOMORROW!

- **Oral Session**: Thursday 10:45, Lehar 1-4
- **Poster Session**: Thursday 11:30, Hall C 4-9 #407
- **Paper**: [arXiv:2402.12535](https://arxiv.org/abs/2402.12535)
- **GitHub**: https://github.com/Graph-COM/HEPT



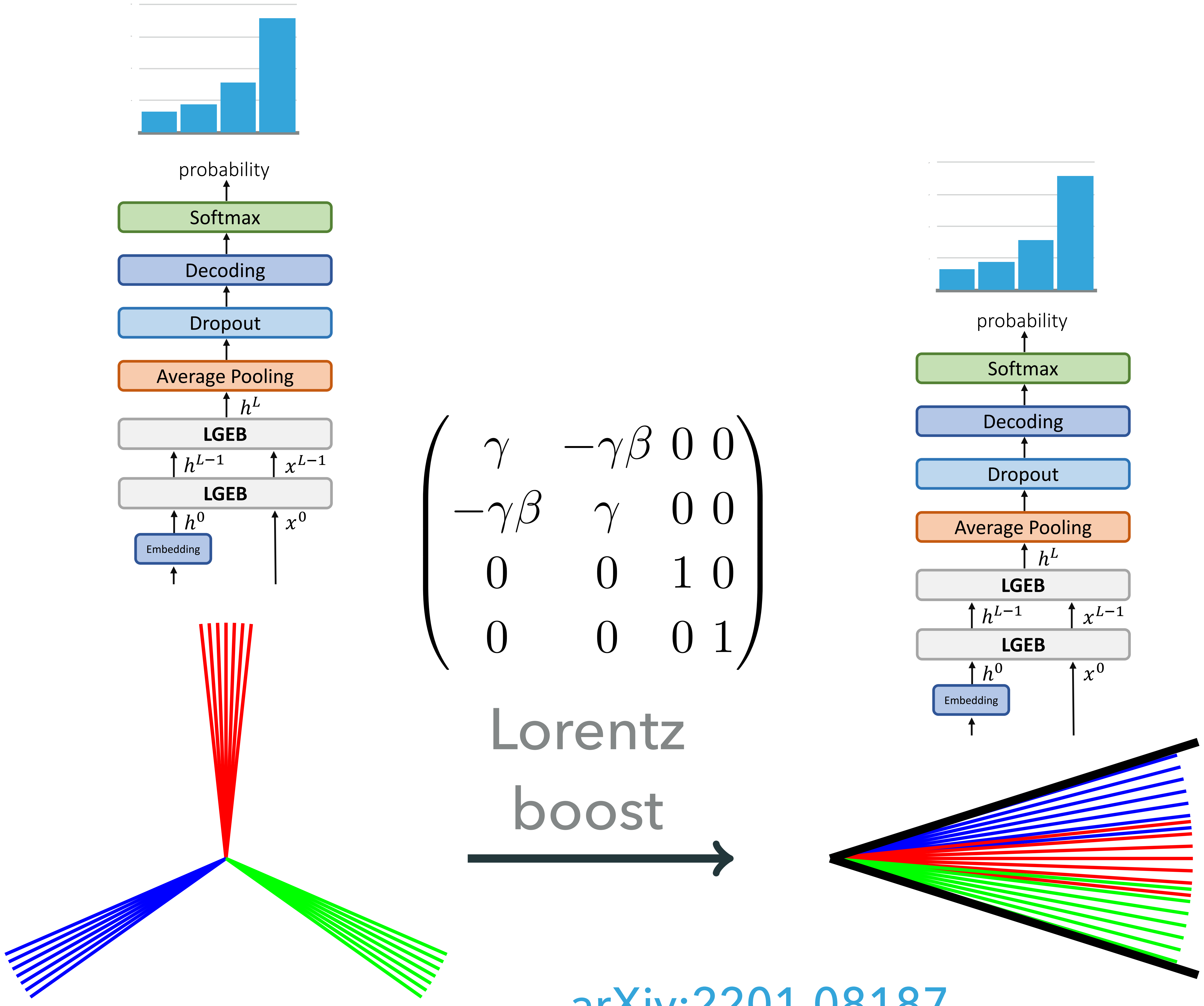


Siqi Miao[1] Zhiyuan Lu[2] Mia Liu[3] Javier Duarte[4] Pan Li[1]

Georgia Tech
北京郵電大學 Beijing University of Posts and Telecommunications
PURDUE UNIVERSITY
UC San Diego

HOW DO WE ENFORCE SYMMETRY?

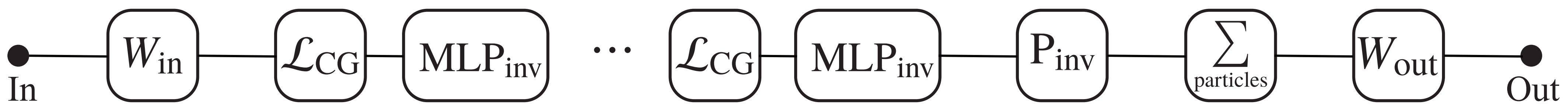
- Lorentz symmetry: physics is the same no matter which reference frame we consider
- Lorentz-invariant networks:
  - Boosting all particles into a new reference frame should give the same result
- Lorentz-equivariant networks:
  - Boosting all particles into a new reference frame should give an output that transforms the same way



arXiv:2201.08187

MANY APPROACHES

- Lorentz Group Network [arXiv:2006.04780]: nonlinearity is tensor product followed by Clebsch-Gordan (CG) decomposition



- LorentzNet [arXiv:2201.08187] uses structured message passing based on Lorentz scalars and vectors
- PELICAN [arXiv:2307.16506] builds invariants and covariants based on pairs of inputs

<u>Invariants</u> (output for PELICAN's "scalar form")

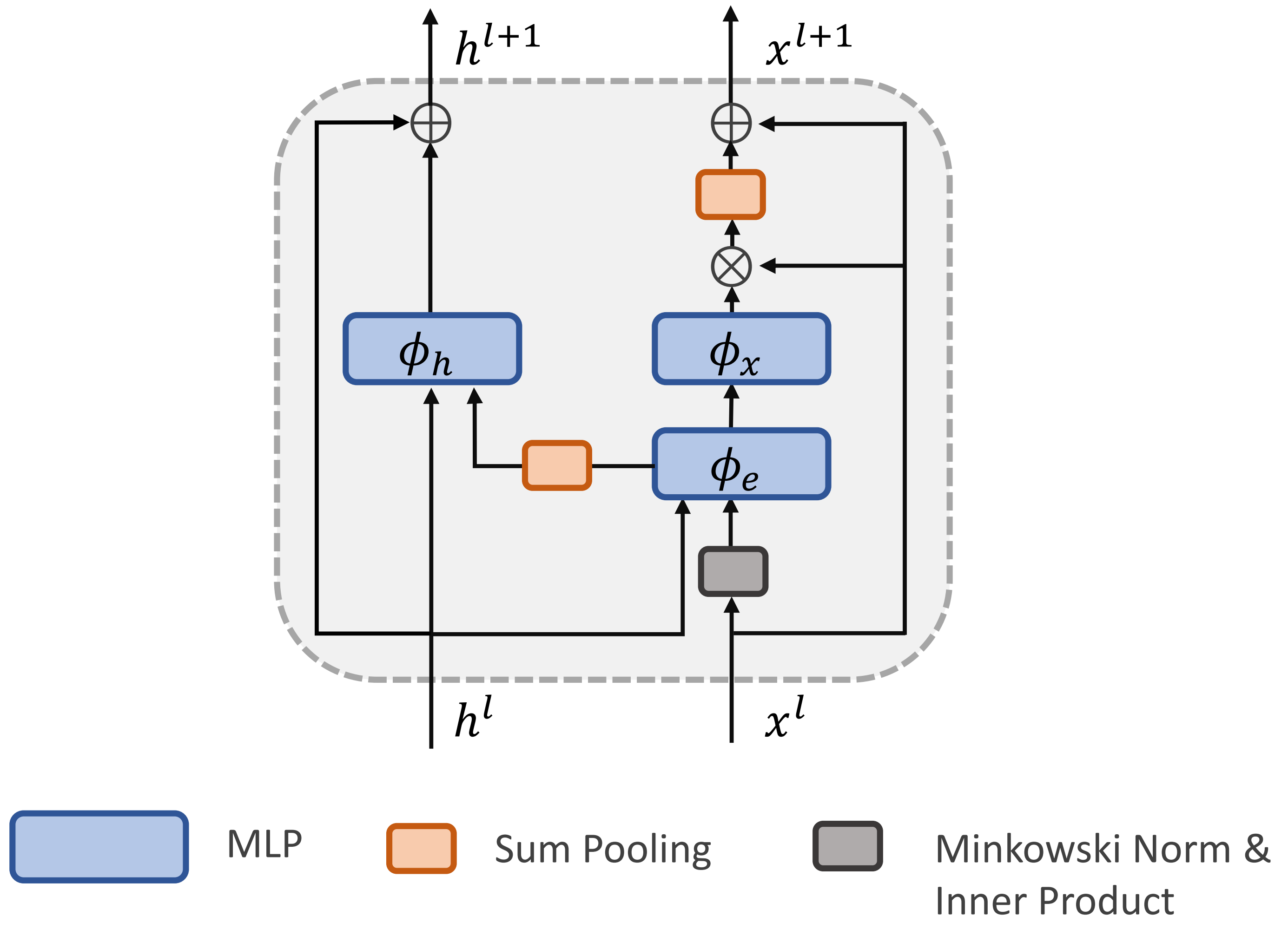
$$I_k\left(p_1,\ldots,p_N\right) = I_k\left(\{d_{ij}\}_{i,j}\right)\ ,\quad d_{ij} \equiv p_i^{\mu} p_{j,\mu}$$

Lorentz-invariant

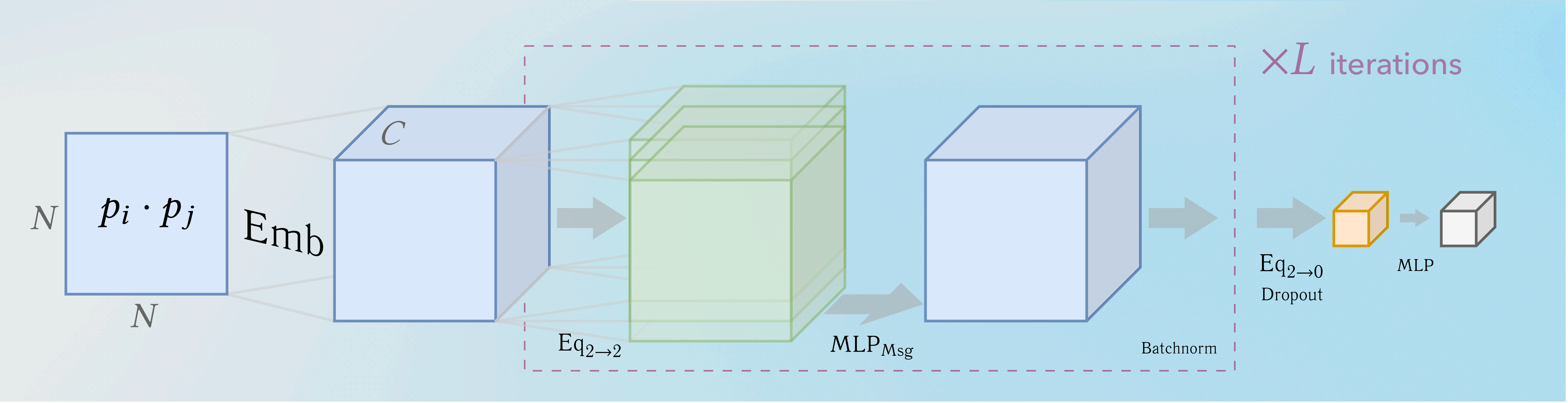
<u>Covariants</u> (output for PELICAN's "vector form")

$$F^{\mu} = \sum_k I_k\left(p_1,\ldots,p_N\right) p_k^{\mu}$$

Lorentz-covariant

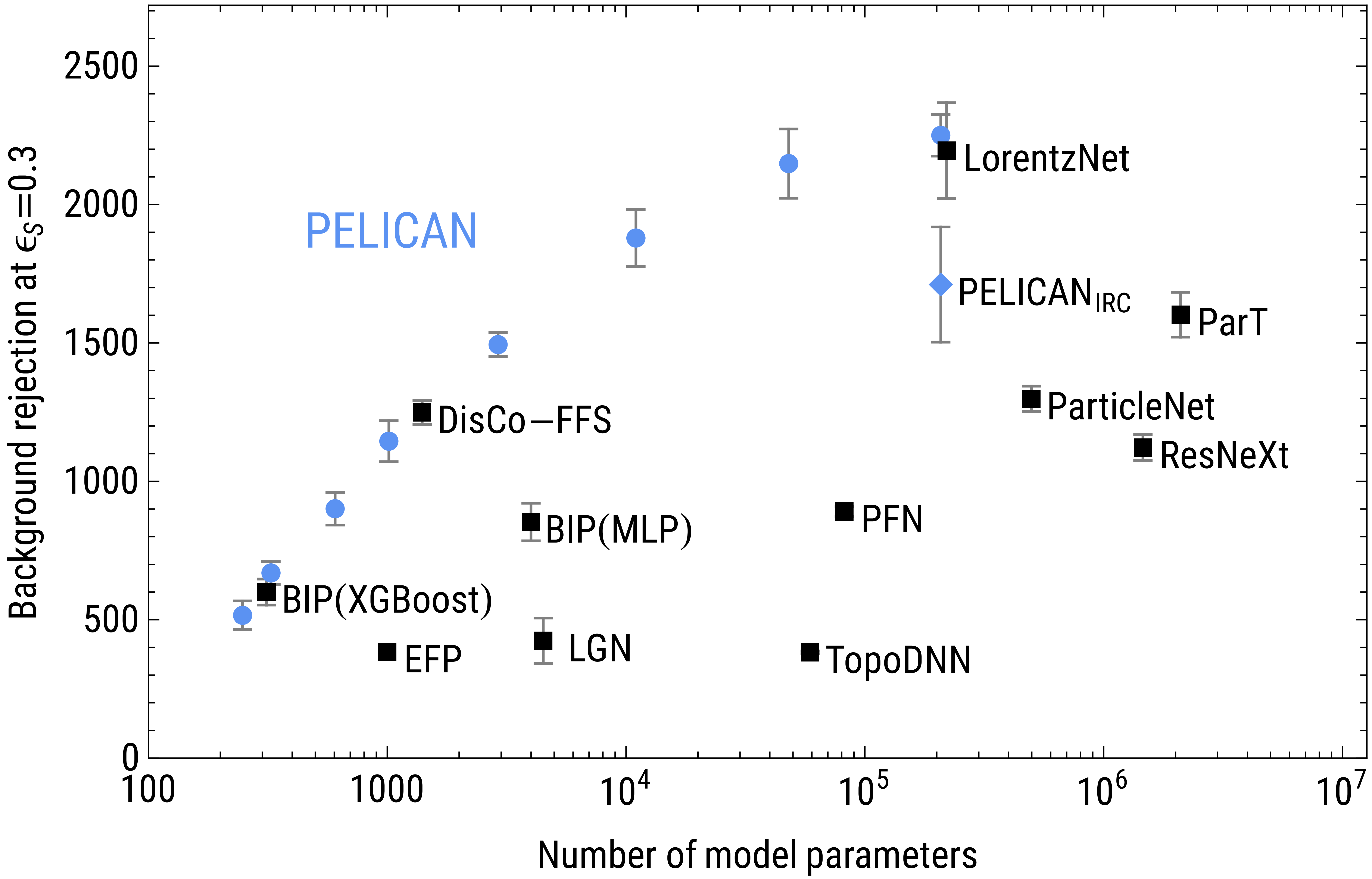


Lorentz Group Equivariant Block (LGEB)

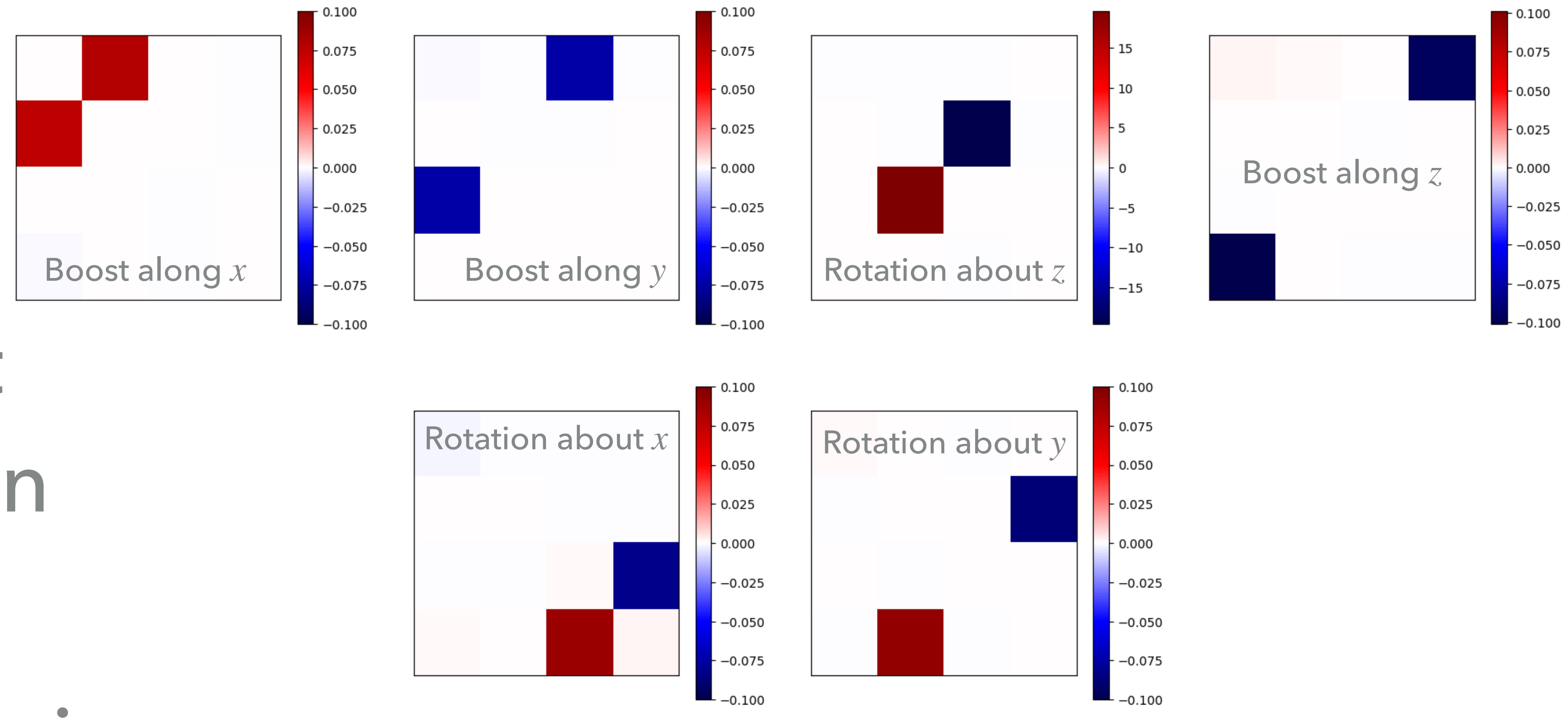


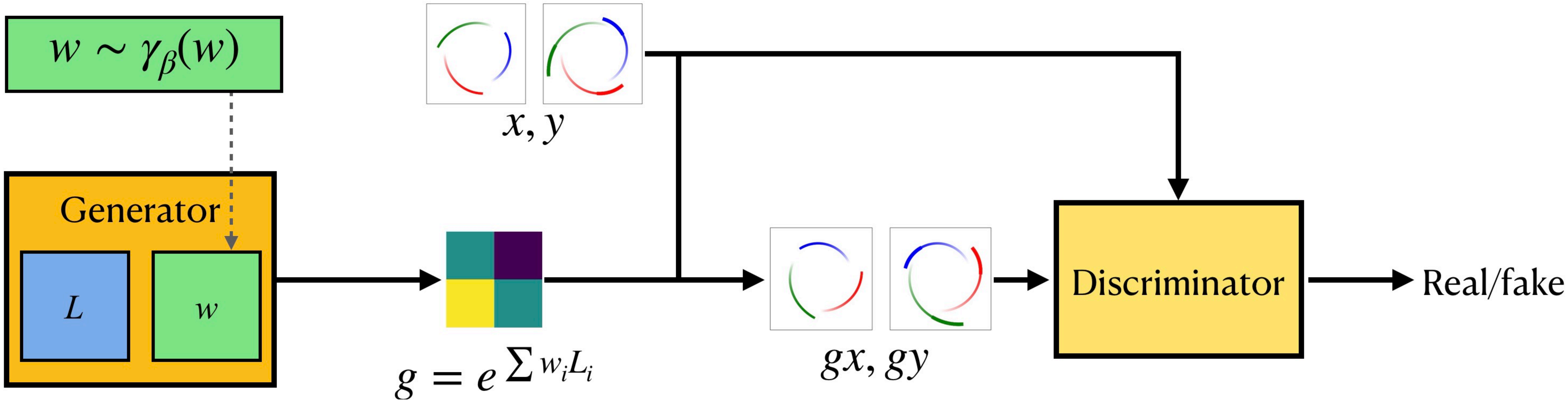
PERFORMANCE AND SCALING

- PELICAN traces Pareto optimal boundary of performance and model complexity



- Symmetries are fundamental principles in particle physics
- LieGAN learns a continuous Lie group that preserves the original data distribution; can discover symmetries present in the data!
- Discovers approximate $SO(1, 3)^+$ symmetry in particle physics dataset





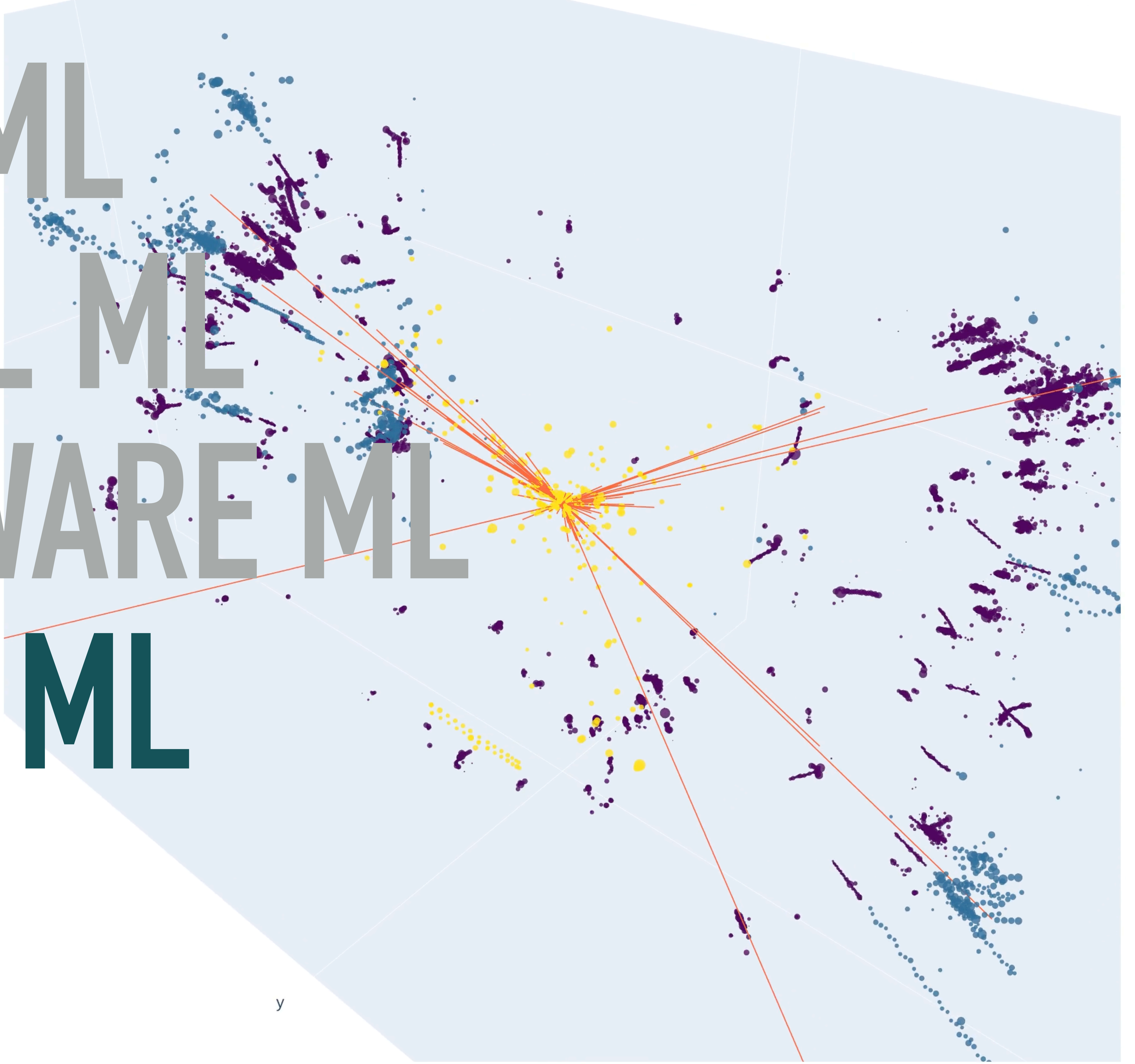
ULTRAFAST ML

MULTIMODAL ML

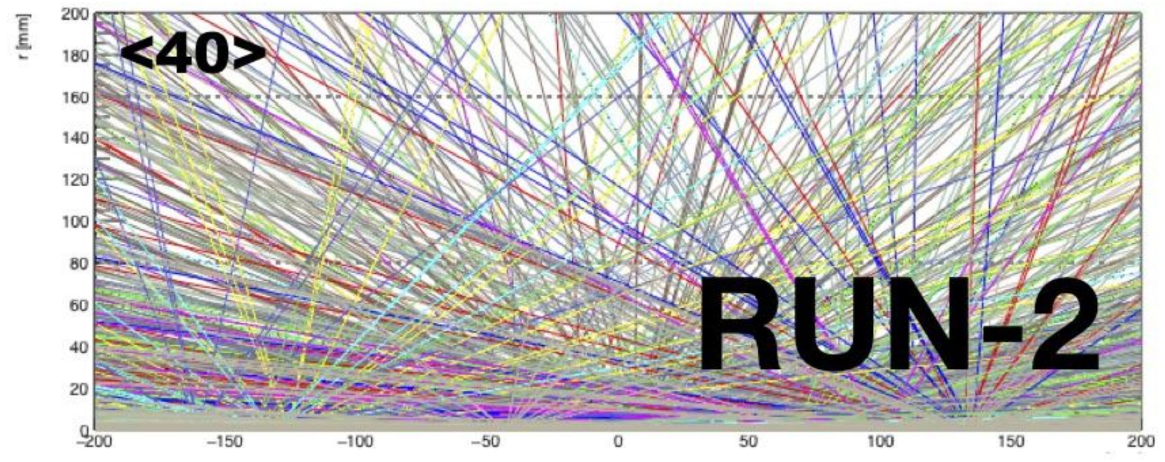
PHYSICS-AWARE ML

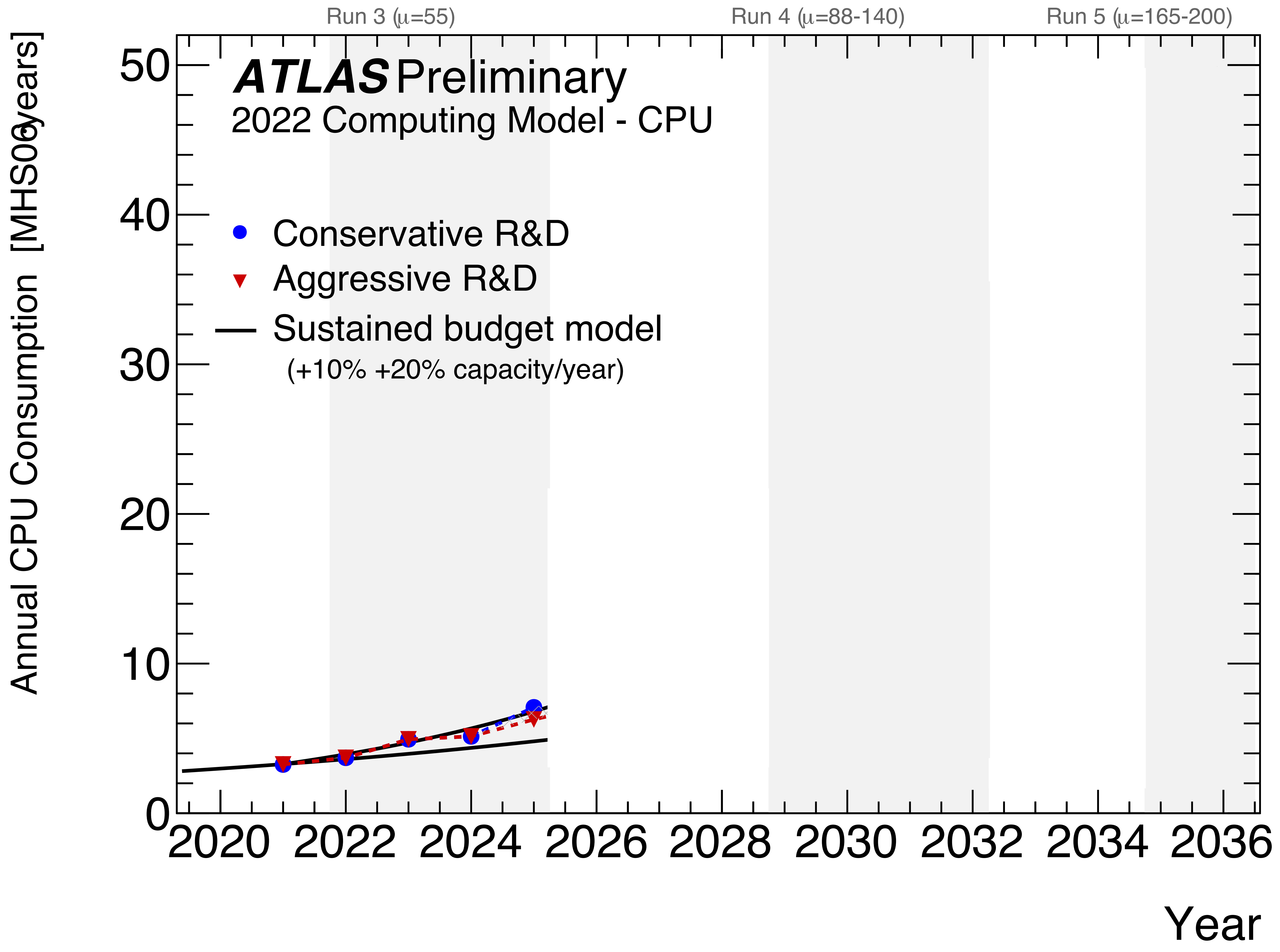
GENERATIVE ML

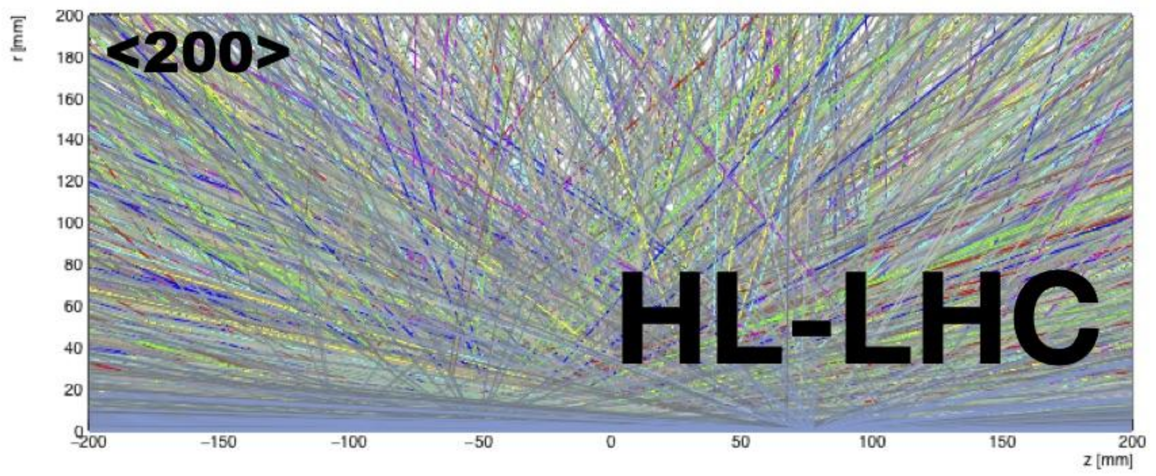
OUTLOOK



FUTURE COMPUTING NEEDS



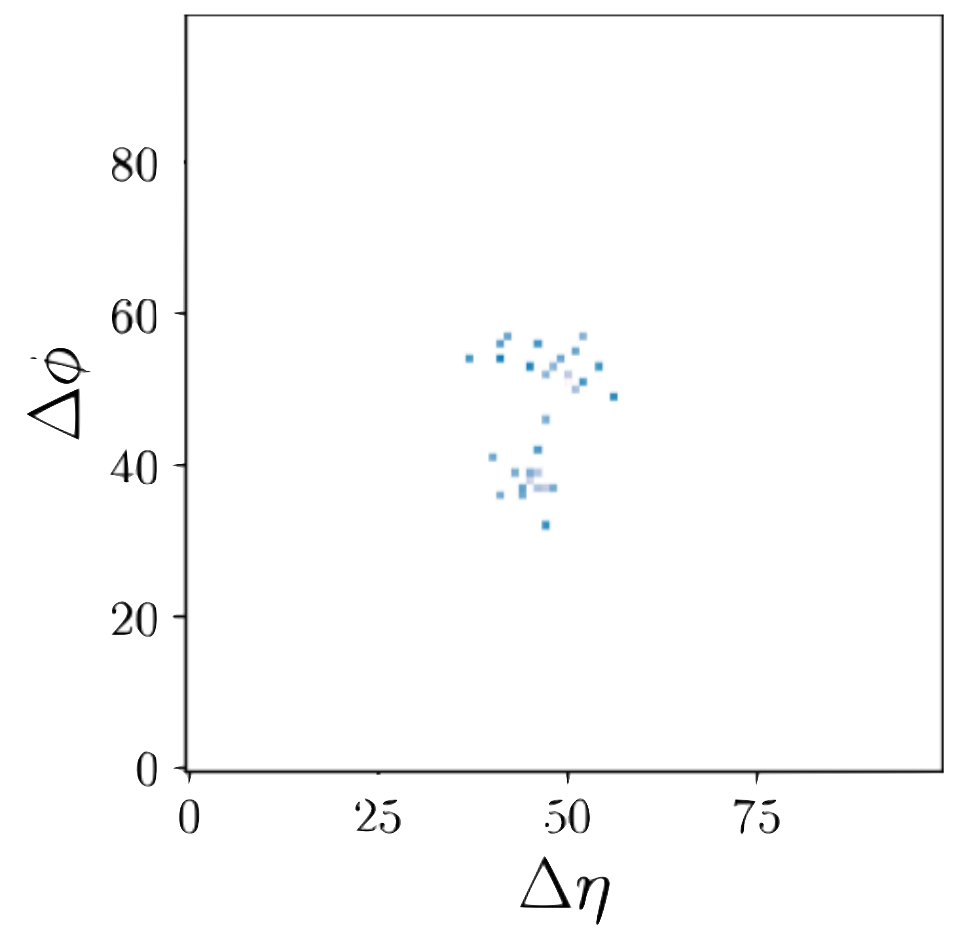


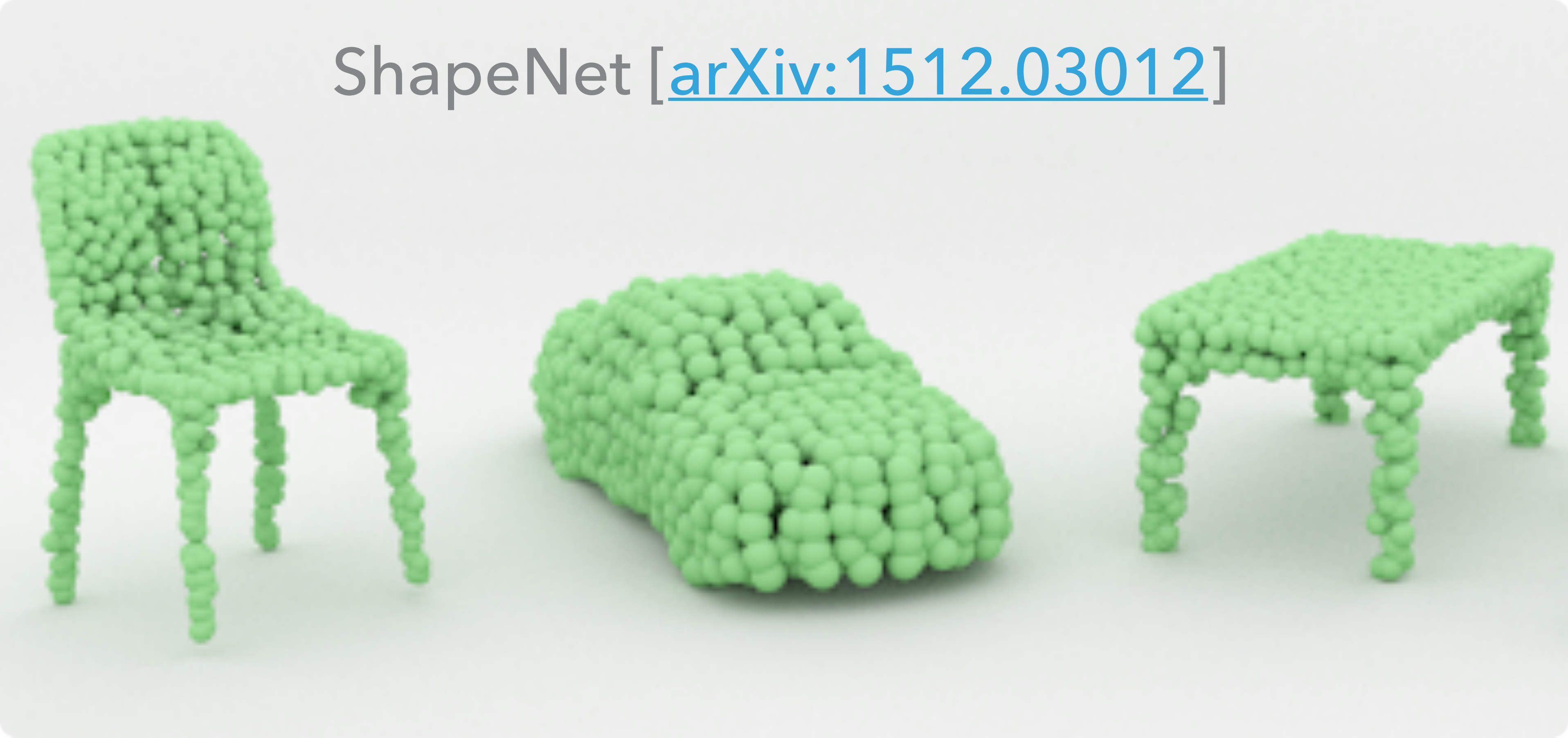


- Simulation is a key driver of CPU needs for the HL-LHC
- ML can be used to “short cut” simulation

JETNET: A BENCHMARK DATASET

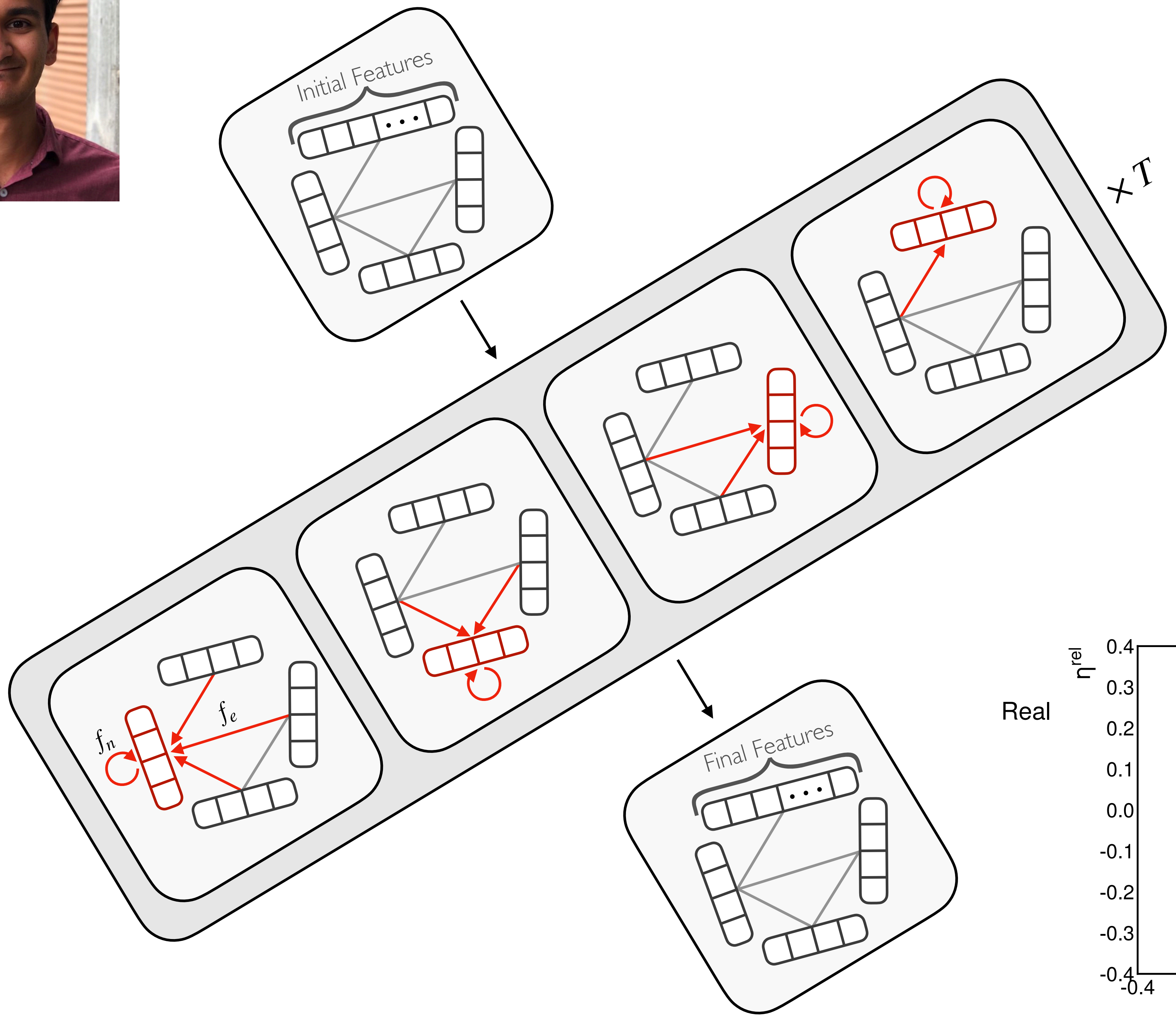
- Public dataset for benchmarking generative models focusing on "particle cloud" representations instead of image representations
  - Similar idea to ShapeNet [arXiv:1512.03012]
  - Consists of up to 150 particles per jet with 3 features: $(p_T^{\mathrm{rel}}, \eta^{\mathrm{rel}}, \phi^{\mathrm{rel}})$
  - Available on Zenodo [doi:10.5281/zenodo.4834875]
- Part of FAIR4HEP [fair4hep.github.io]

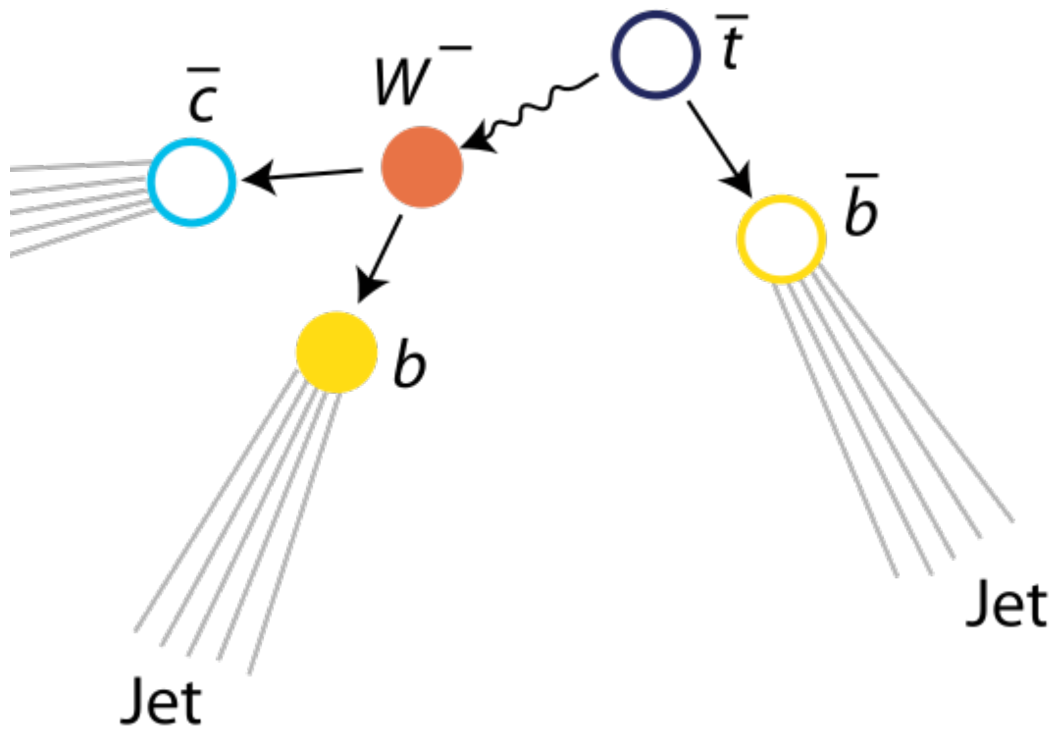


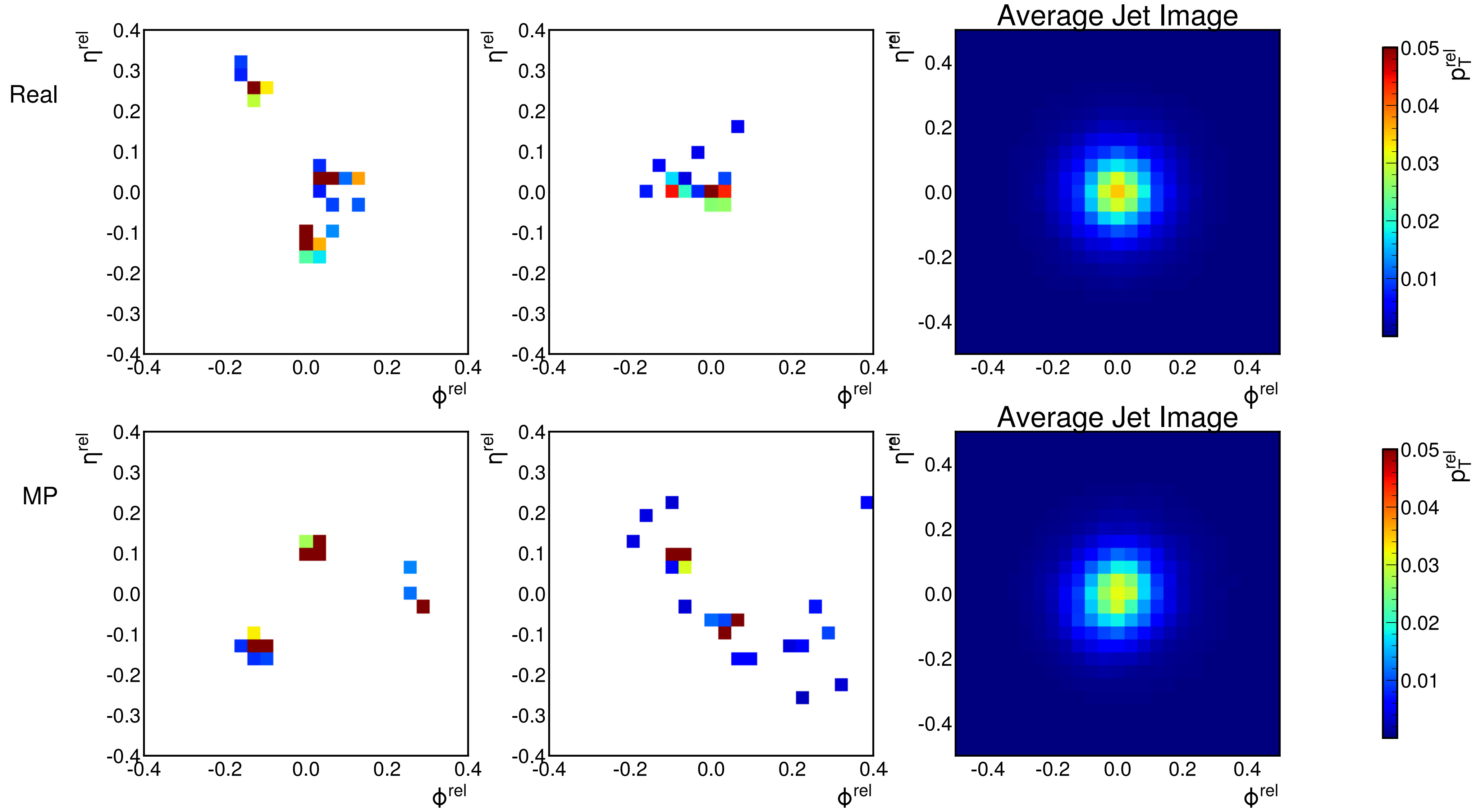


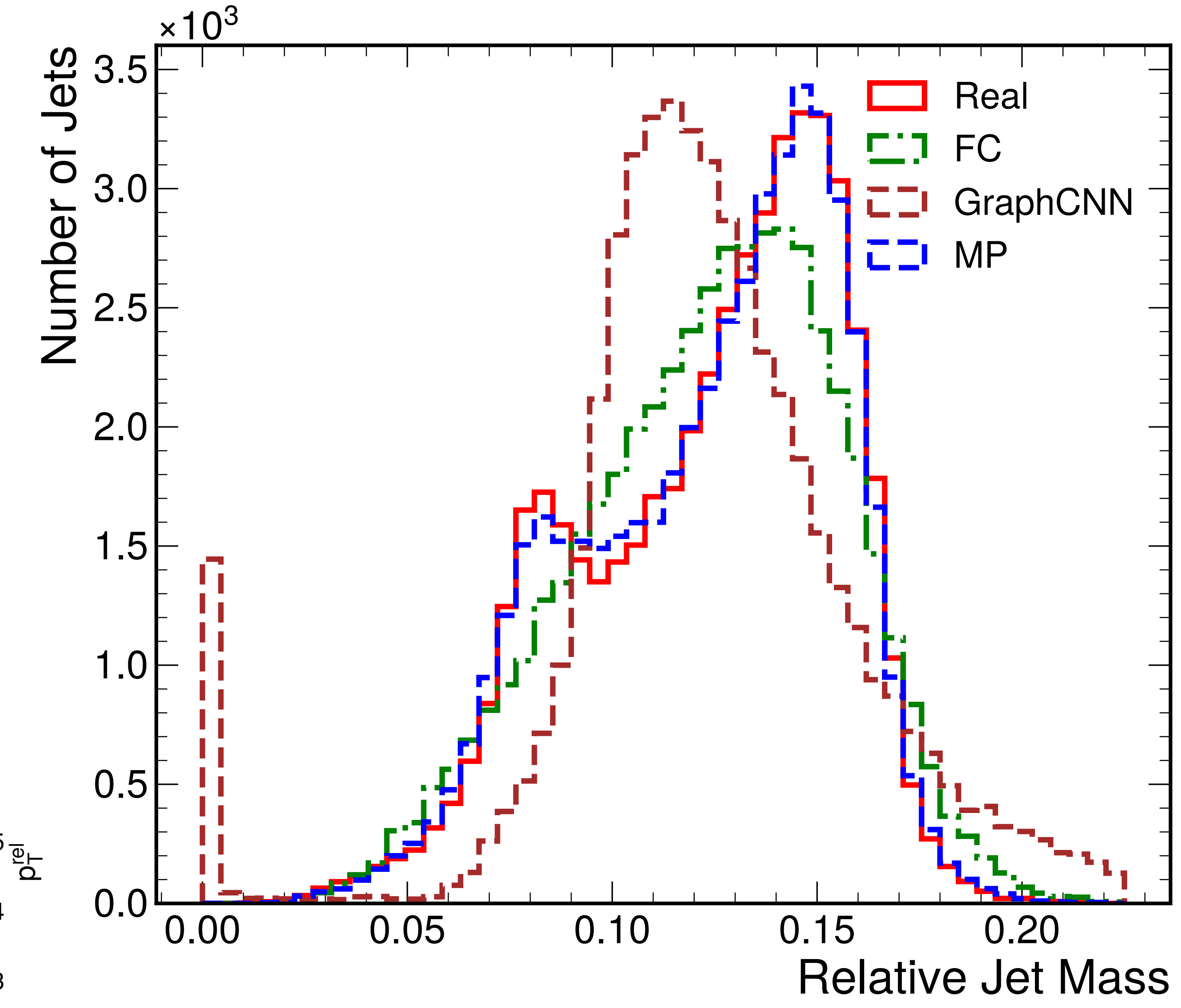
MESSAGE-PASSING GAN

NeurIPS 34, 23858 (2021)





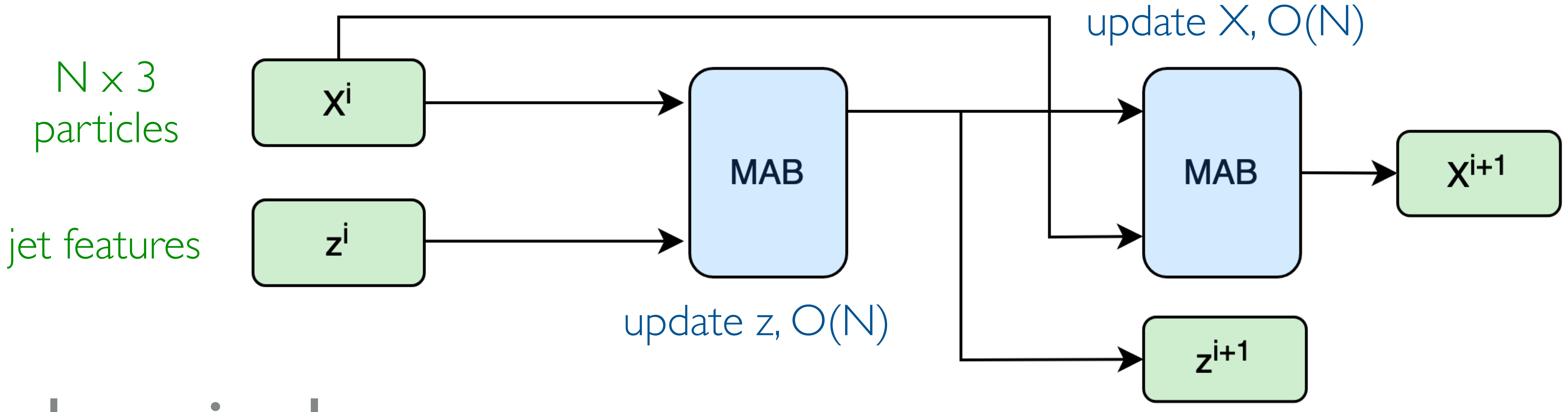




- Message-passing GAN can generate realistic particle jets
- Outperforms existing point cloud GANs

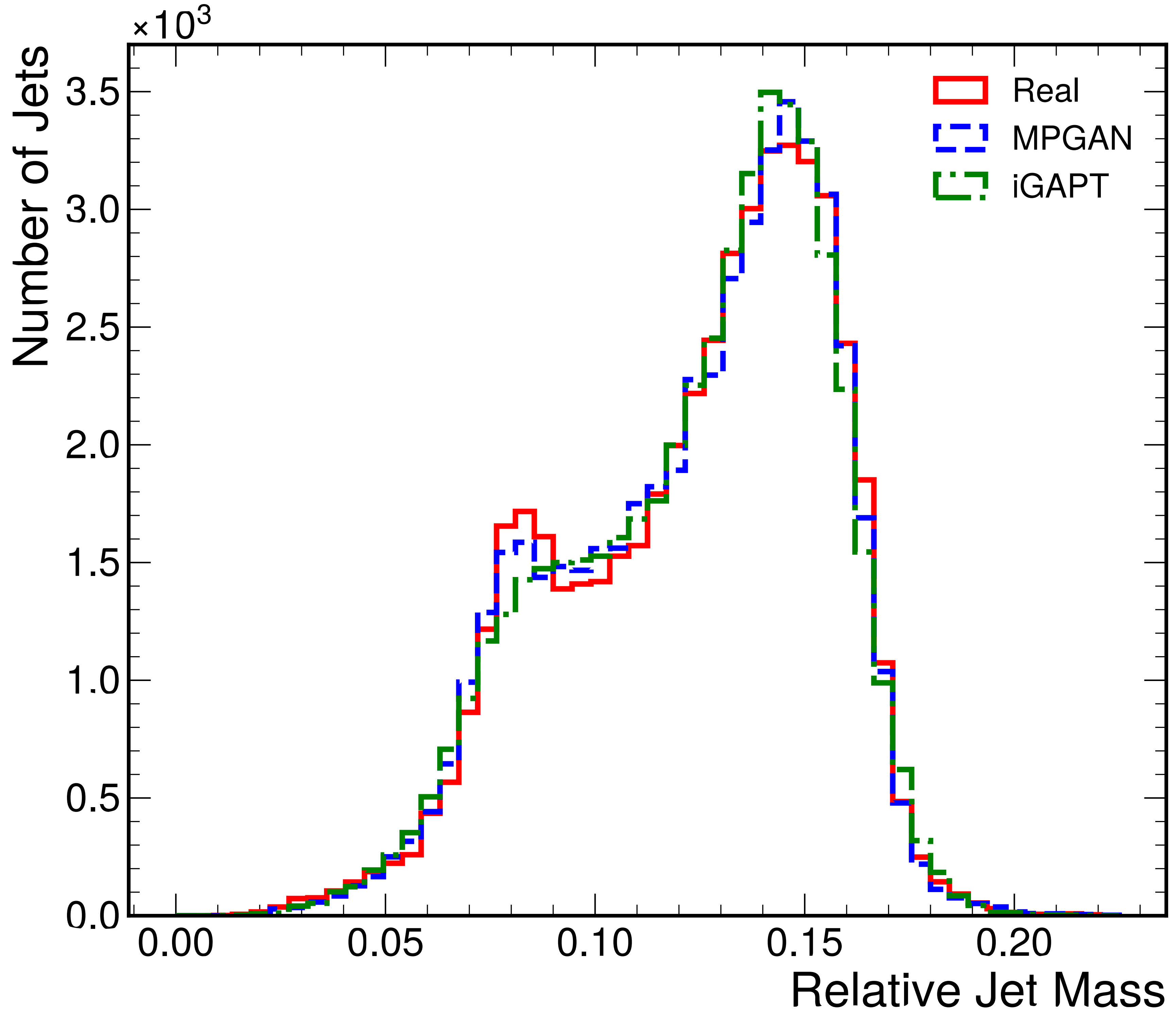
INDUCED GRAPH ATTENTION & METRICS

- Induced graph attention particle transformer (iGAPT) more efficient than MPGAN, but how to compare them?



- Metrics developed to comprehensively evaluate generative AI models in HEP
  - **Fréchet Physics Distance**: Inspired by Fréchet Inception Distance with physics-based features

	FPD $\times 10^3$	Inference time (μs) per jet
Truth	0.08 ± 0.03	
MPGAN	$\mathbf{0.30 \pm 0.06}$	41
GAPT	0.66 ± 0.09	9

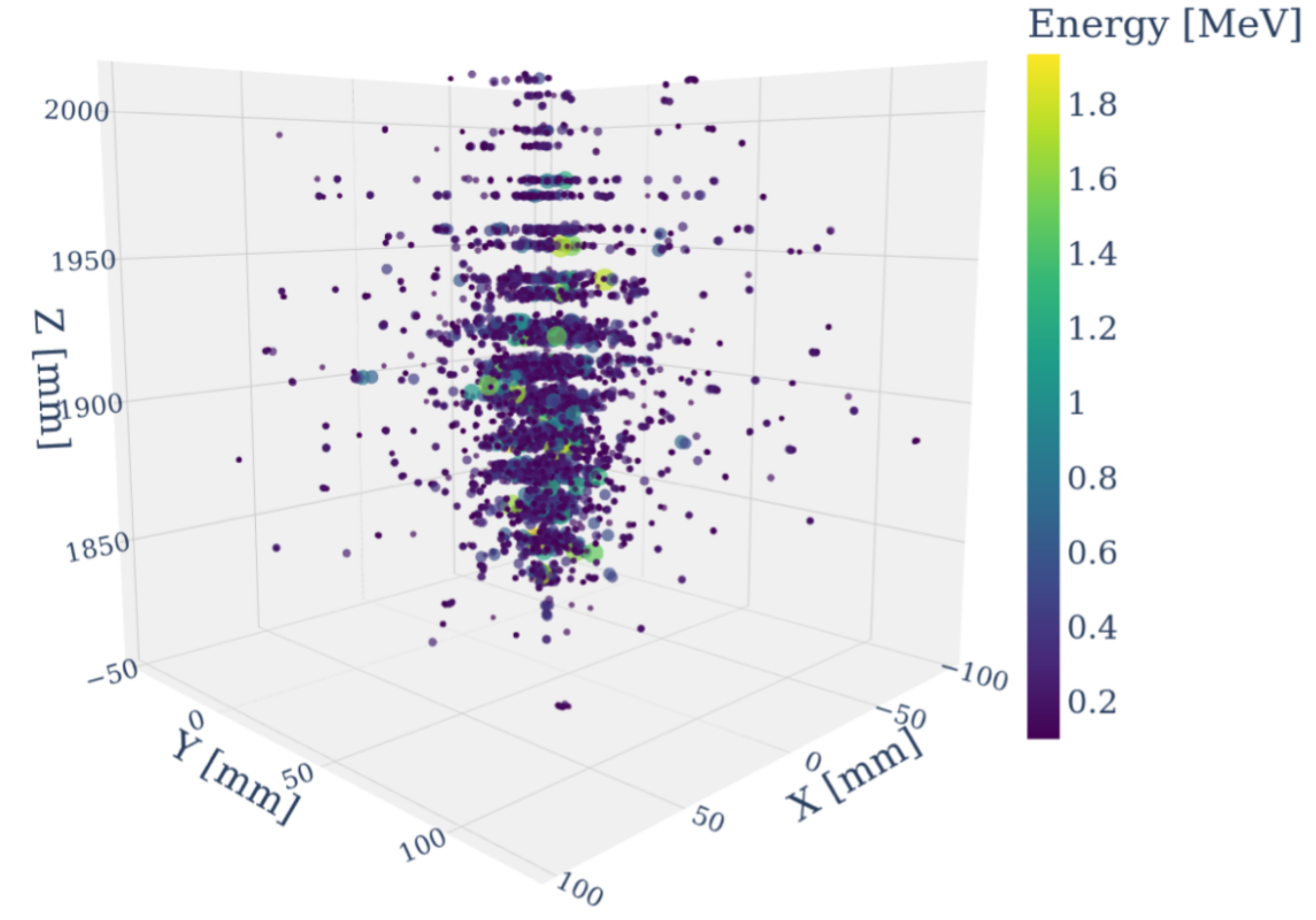


IMPACT OF JETNET

- Sparked significant R&D into equivariant models, **diffusion models**, and more!

CaloClouds: Fast Geometry-Independent Highly-Granular Calorimeter Simulation

Erik Buhmann[1], Sascha Diefenbacher[1,2], Engin Eren[3], Frank Gaede[3,4], Gregor Kasieczka[1,4], Anatolii Korol[3*], William Korcari[1], Katja Krüger[3], Peter McKeown[3]



Pay Attention to Mean-Fields during Particle Cloud Generation

Benno Käch
Deutsches Elektronen-Synchrotron DESY
Germany
benno.kaech@desy.de

Isabell-A. Melzer-Pellmann
Deutsches Elektronen-Synchrotron DESY
Germany
isabell.melzer@desy.de

SciPost Physics — Submission

PC-JeDi: Diffusion for Particle Cloud Generation in High Energy Physics

Matthew Leigh, Debajyoti Sengupta, Guillaume Quétant, John Andrew Raine, Knut Zoch, and Tobias Golling,

University of Geneva

SciPost Physics — Submission

MIT-CTP 5519

EPiC-GAN: Equivariant Point Cloud Generation for Particle Jets

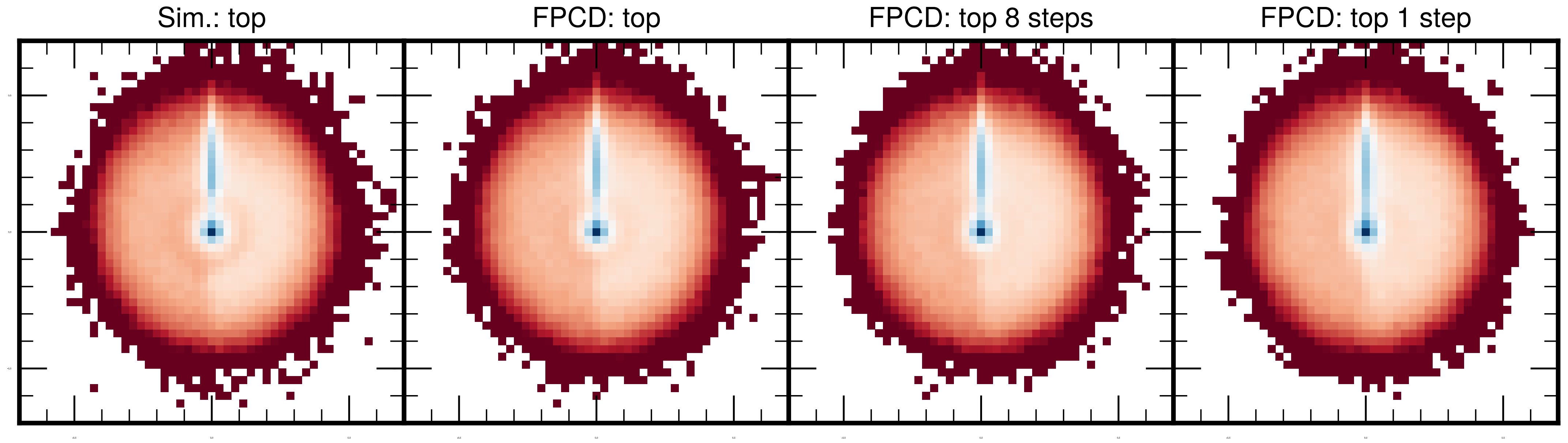
Erik Buhmann[1*], Gregor Kasieczka[1,2] and Jesse Thaler[3,4]

1 Institut für Experimentalphysik, Universität Hamburg, Germany
2 Center for Data and Computing in Natural Sciences (CDCS), Hamburg, Germany
3 Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States
4 The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
* erik.buhmann@uni-hamburg.de

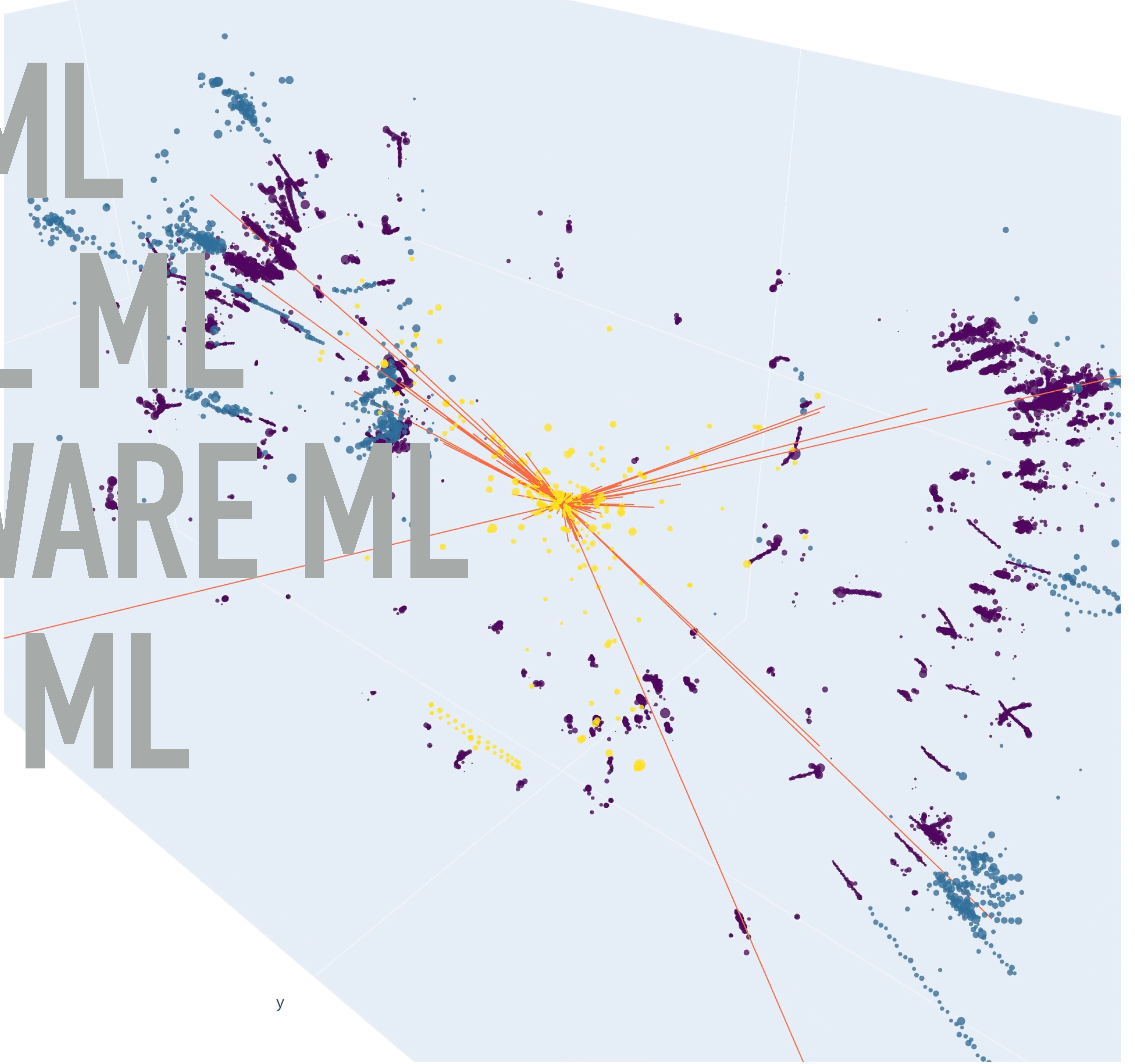
Fast Point Cloud Generation with Diffusion Models in High Energy Physics

Vinicius Mikuni,[1,*] Benjamin Nachman,[2,3,†] and Mariel Pettee[2,‡]

[1] *National Energy Research Scientific Computing Center, Berkeley Lab, Berkeley, CA 94720, USA*
[2] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[3] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*

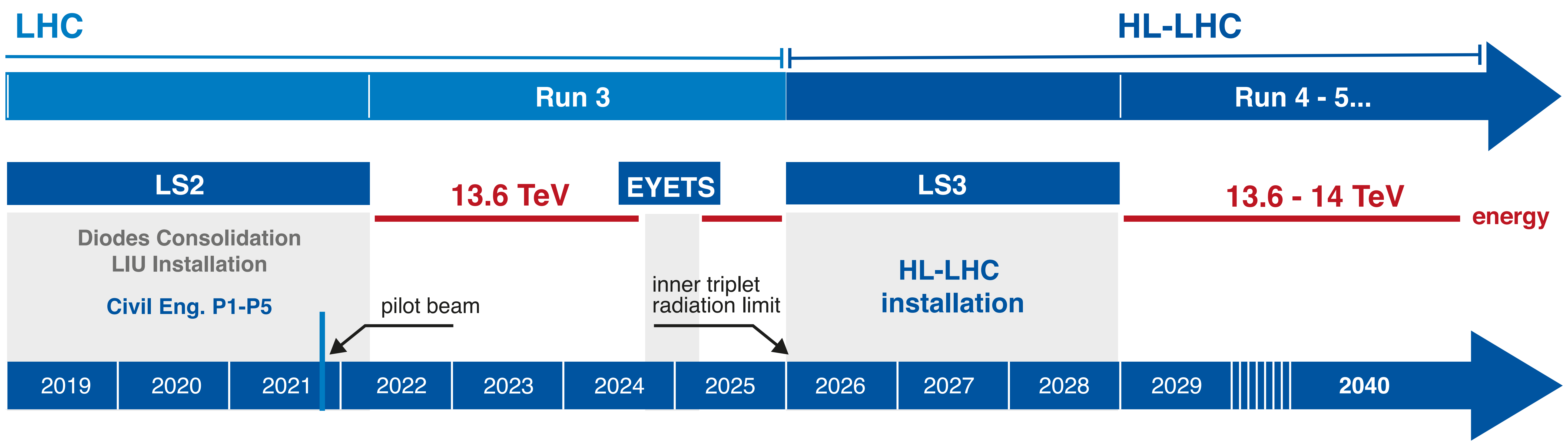


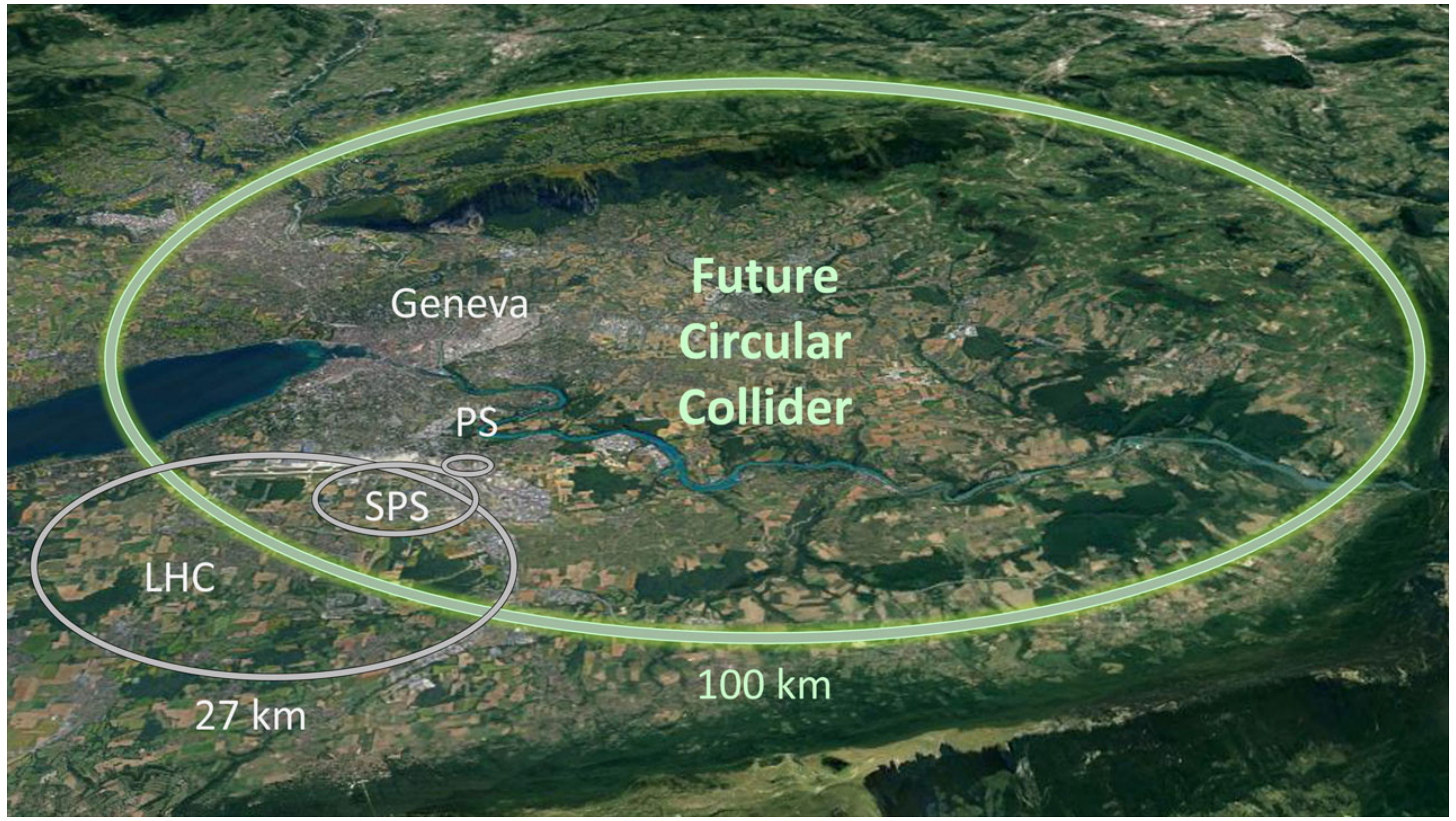
ULTRAFAST ML
MULTIMODAL ML
PHYSICS-AWARE ML
GENERATIVE ML
OUTLOOK

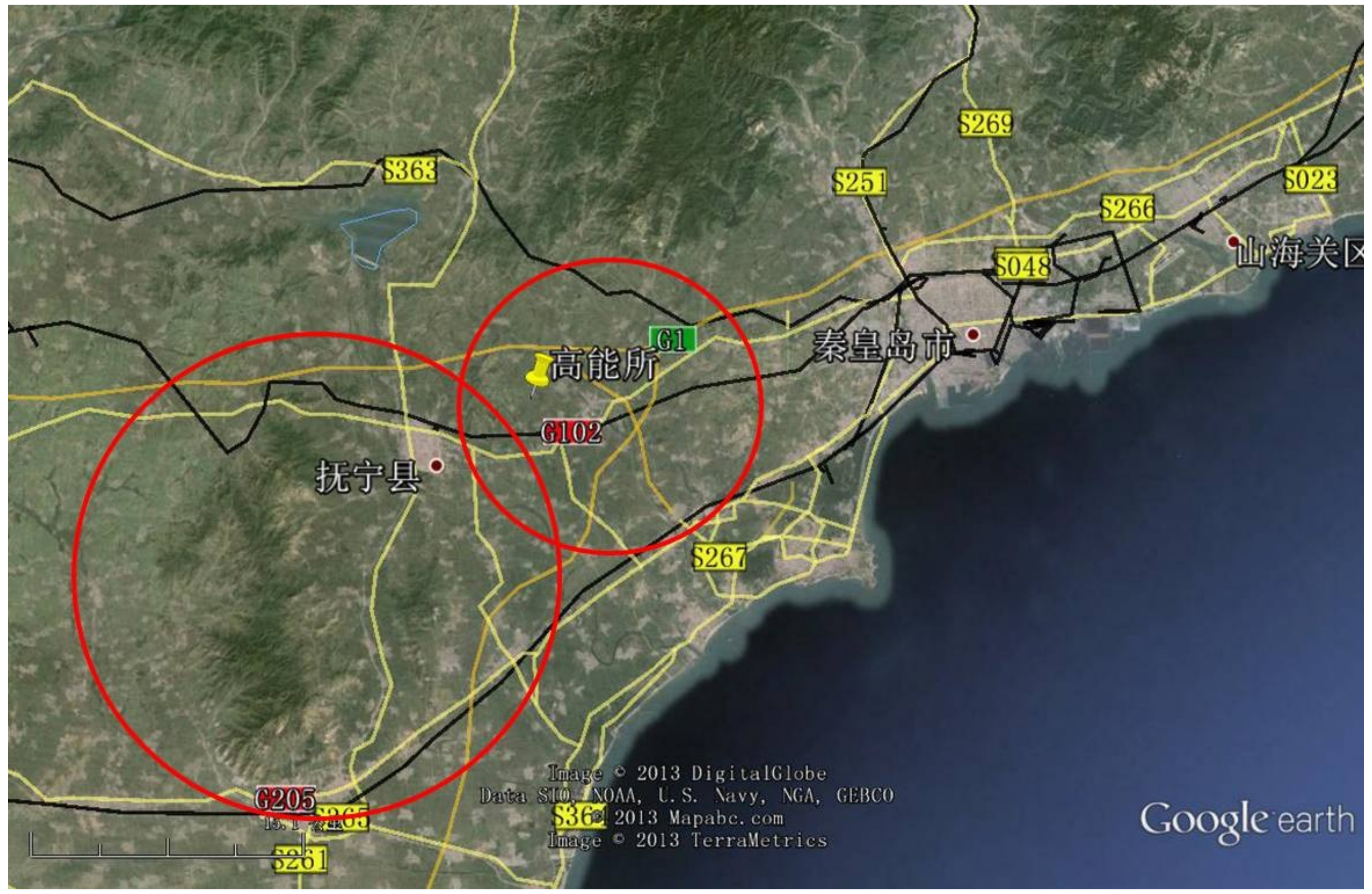


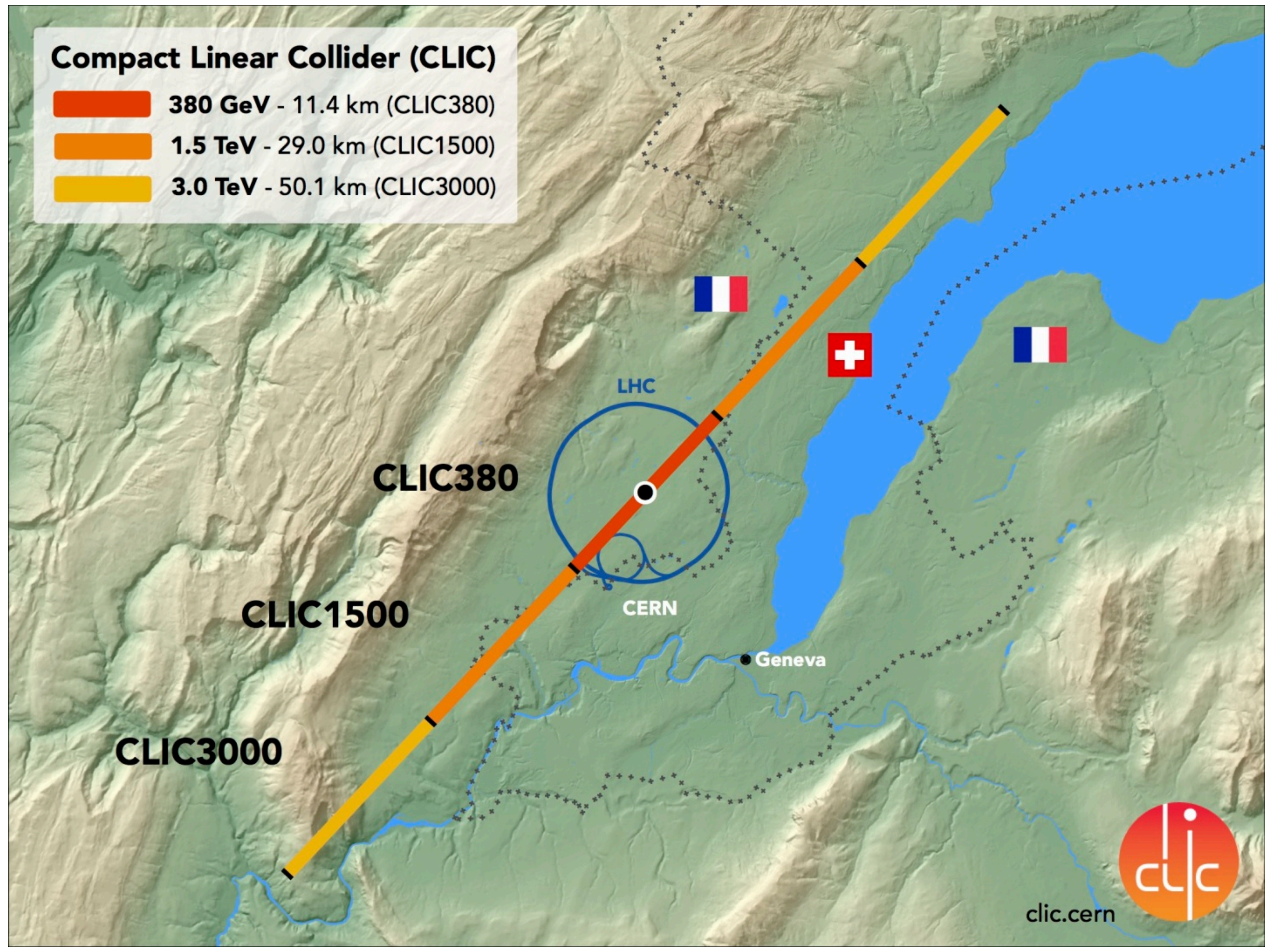
THE FUTURE OF PARTICLE PHYSICS

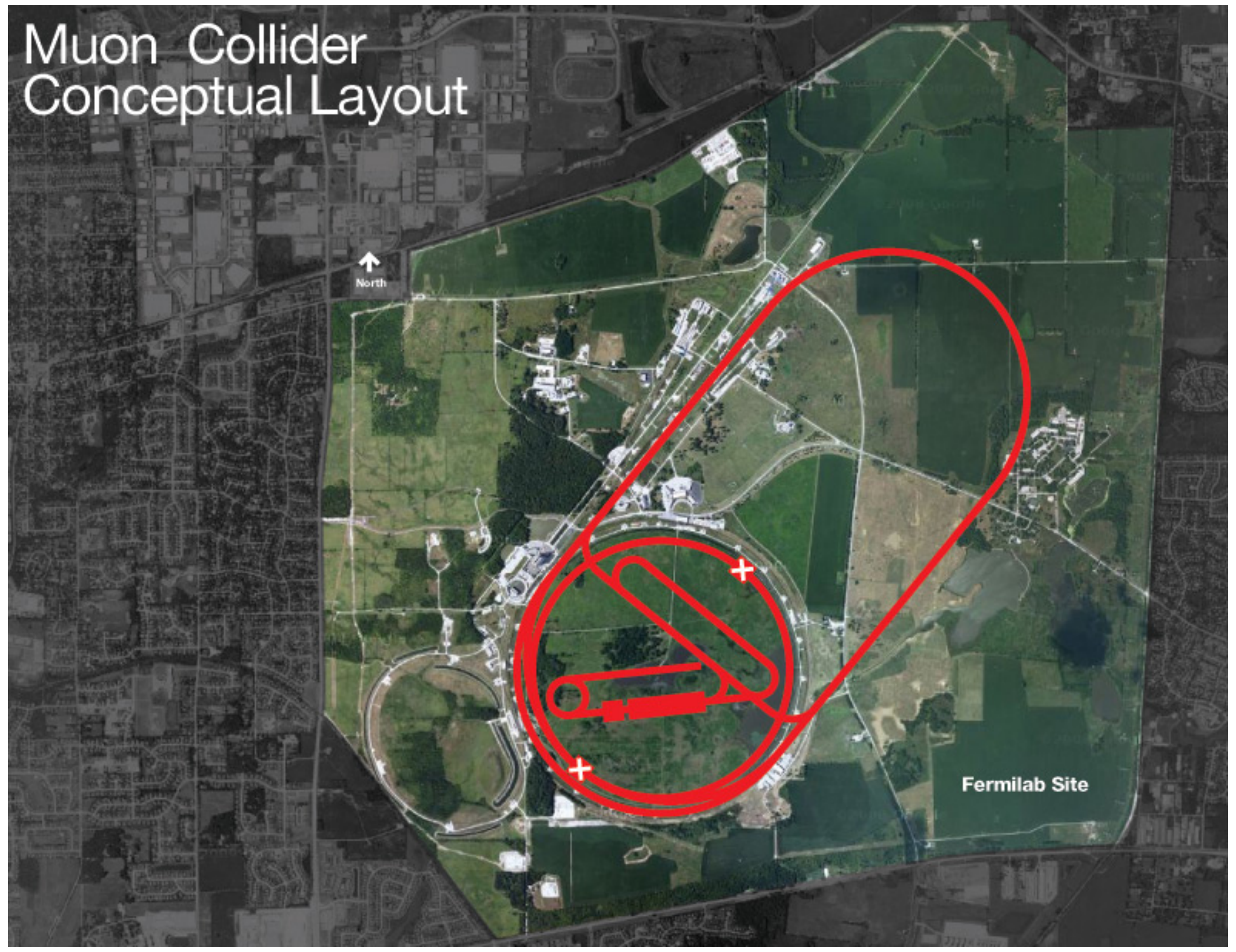
- We have been considering what should come after the HL-LHC…





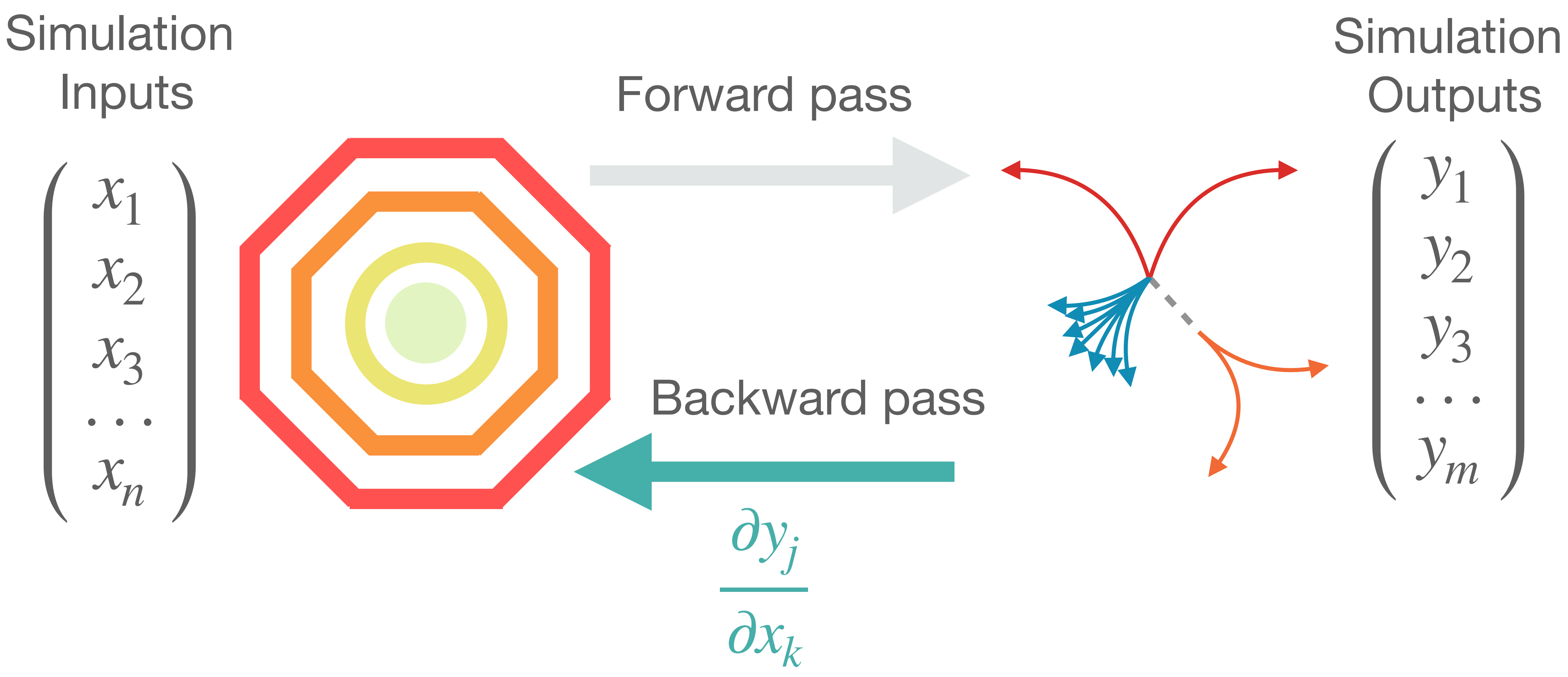


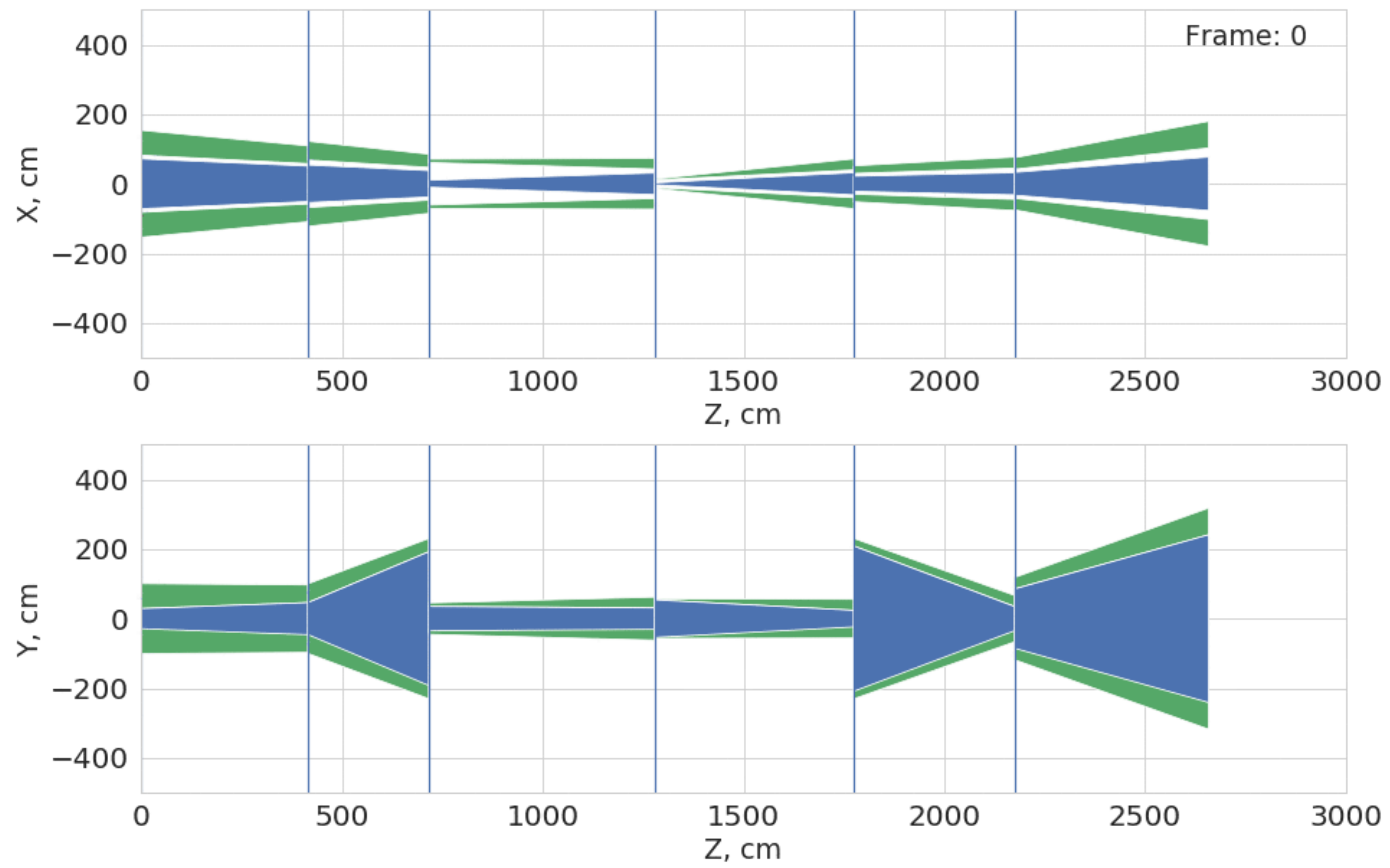




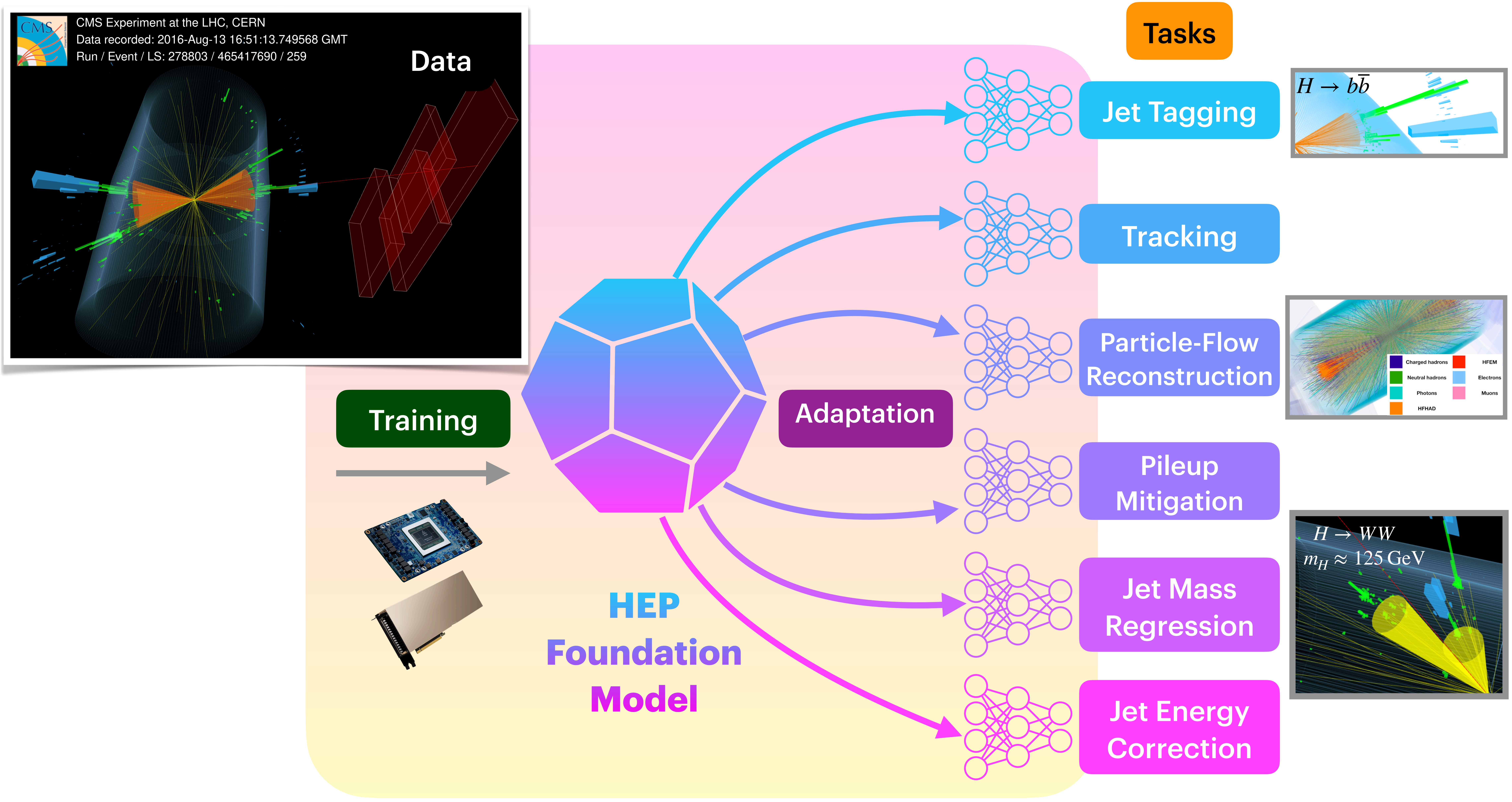
ML TO OPTIMIZE DESIGN DETECTORS

- ML has the promise to help us optimize the design of future colliders and detectors
  - Need all aspects of simulation chain to be implemented with differentiable programming [arXiv:2002.04632]
- Check out Differentiable Almost Everything Workshop



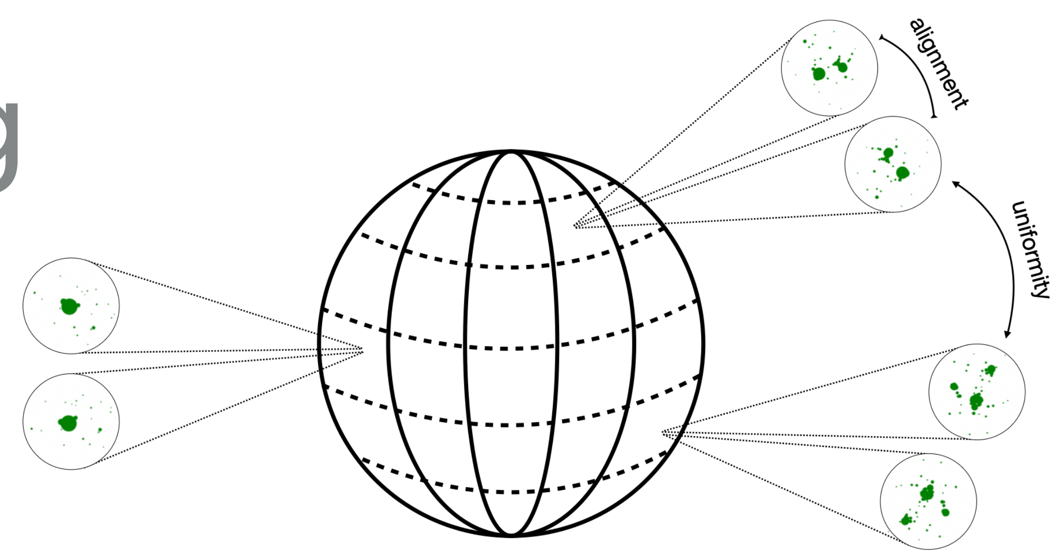


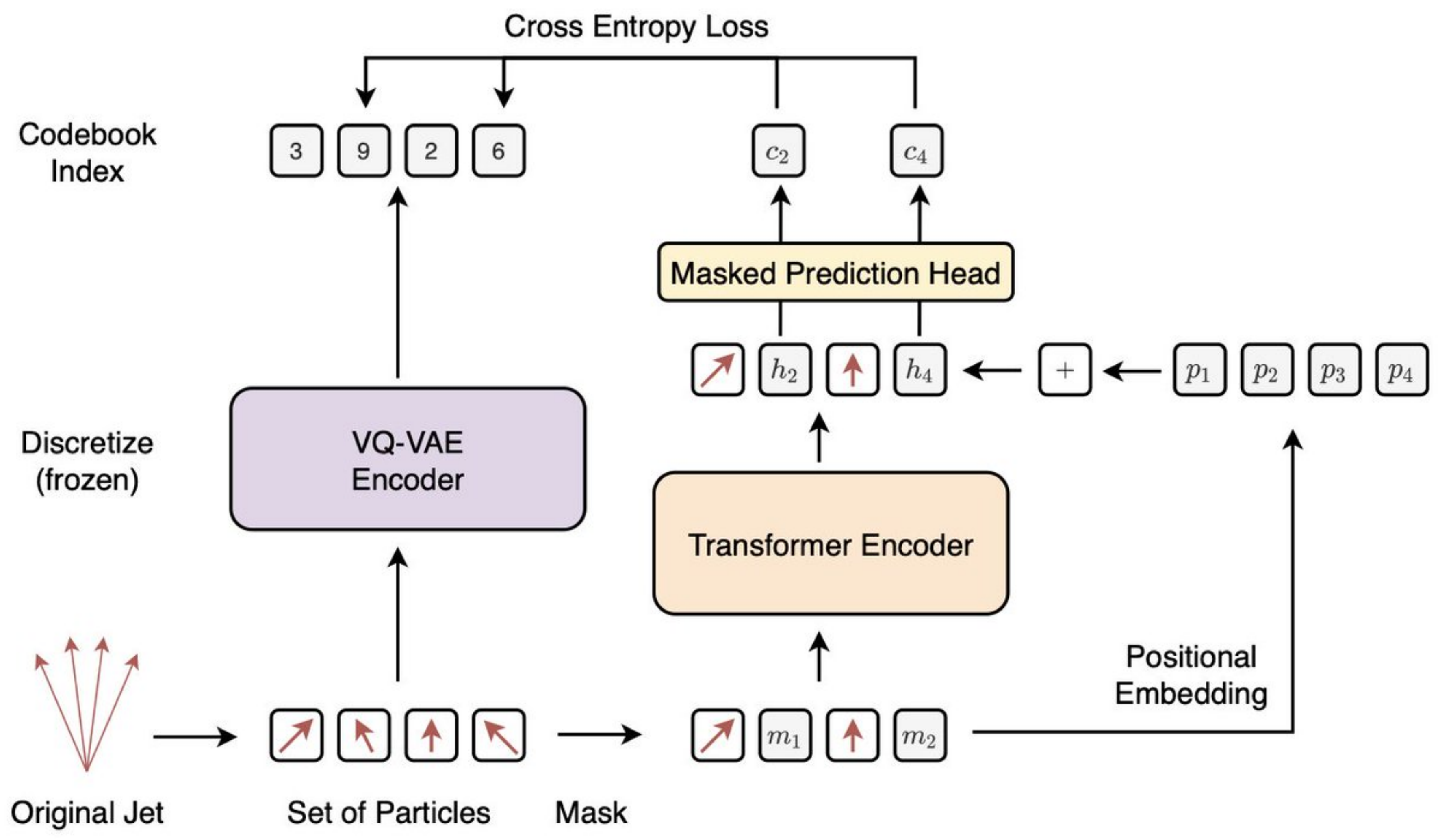
A HEP FOUNDATION MODEL?

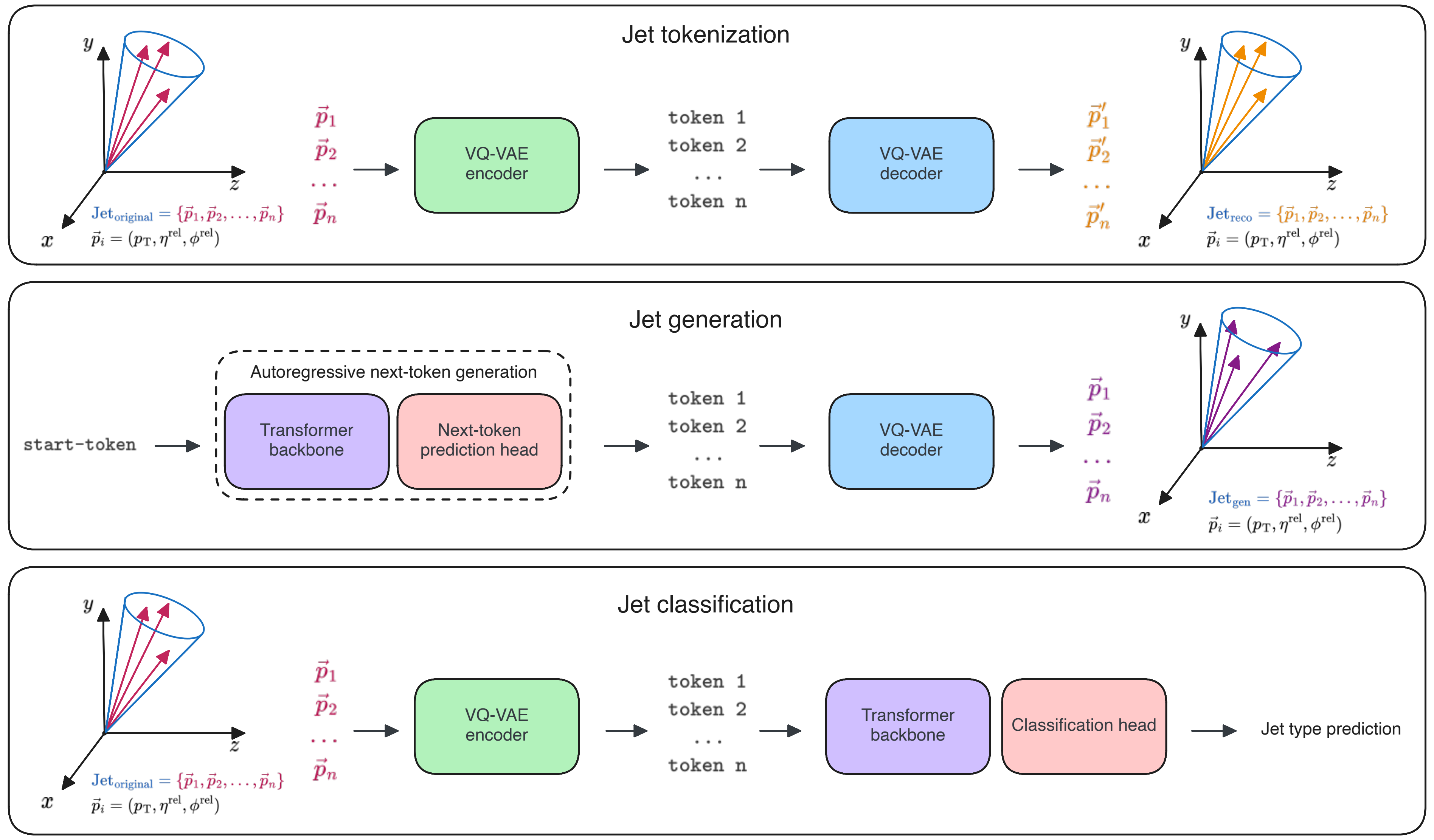


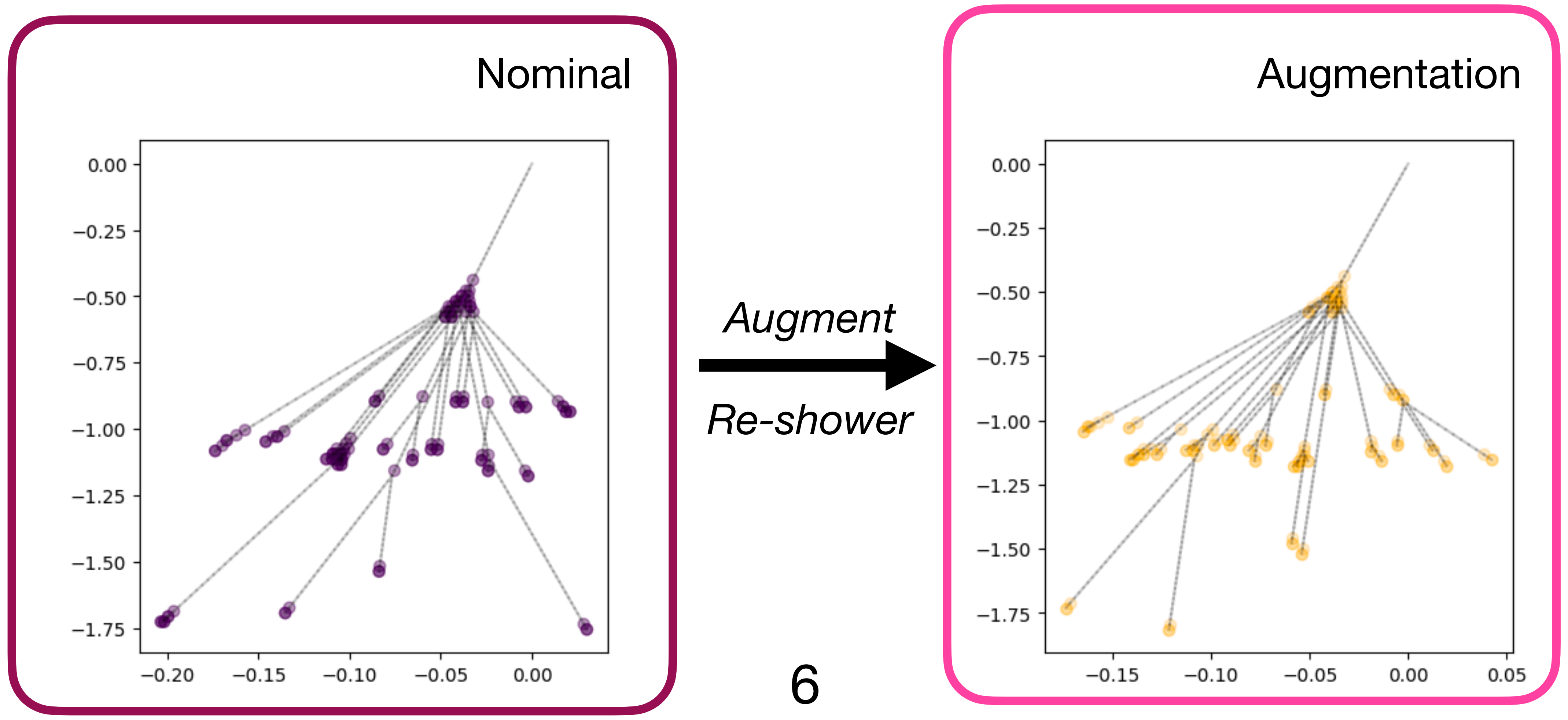
SELF-SUPERVISED LEARNING IN HEP

- Many studies implementing CS- and physics-inspired pre-training strategies
  - JetCLR [arXiv:2108.04253]
  - Masked particle modeling [arXiv:2401.13537]
  - Resimulation [arXiv:2403.07066]
  - GPT [arXiv:2403.05618]









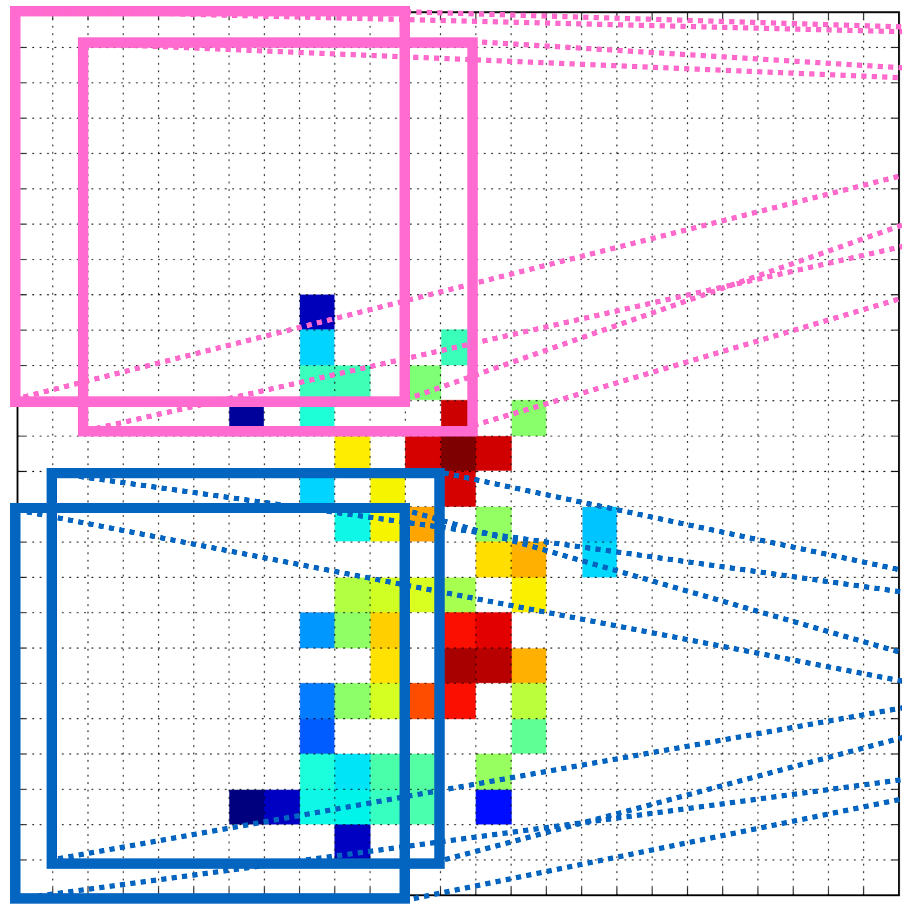
Initial parton fixed, only re-do shower +hadronization
Toy simulation
Credit to: M Kagan L Heinrich

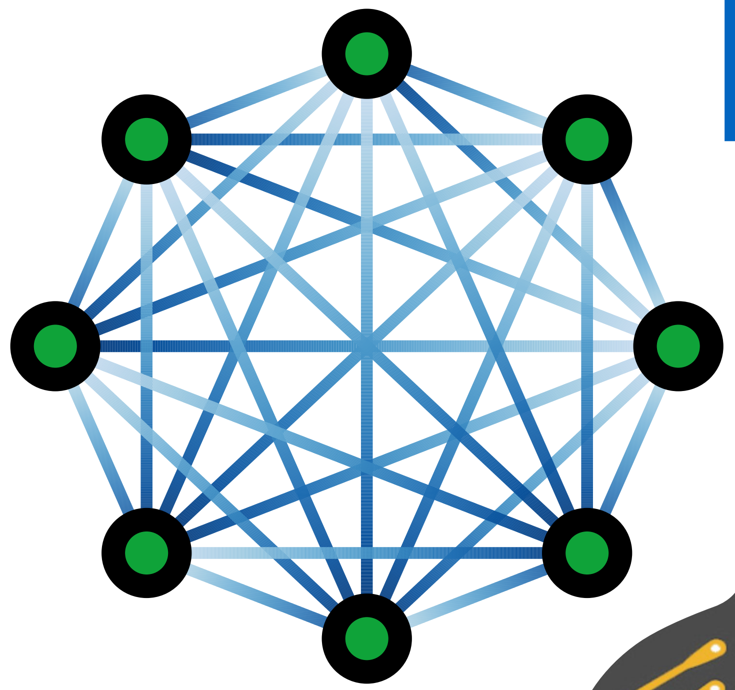
SUMMARY AND OUTLOOK

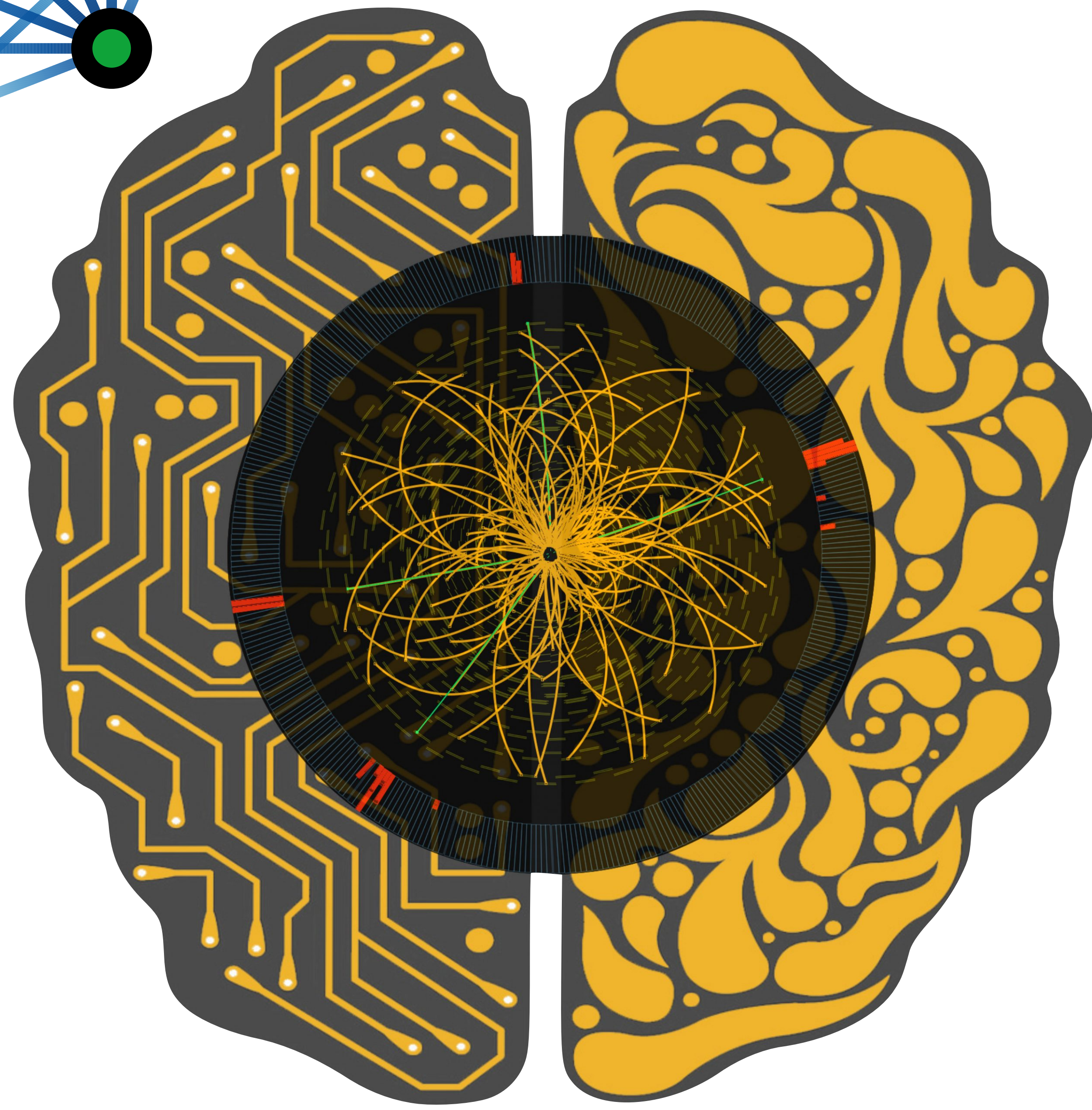
- Dizzying array of ***ML opportunities, innovations, and applications*** in ***particle physics experiments***

- ML can help us solve major challenges for the next generation of particle physics experiments

- So we can (hopefully) get answers to **our big questions**

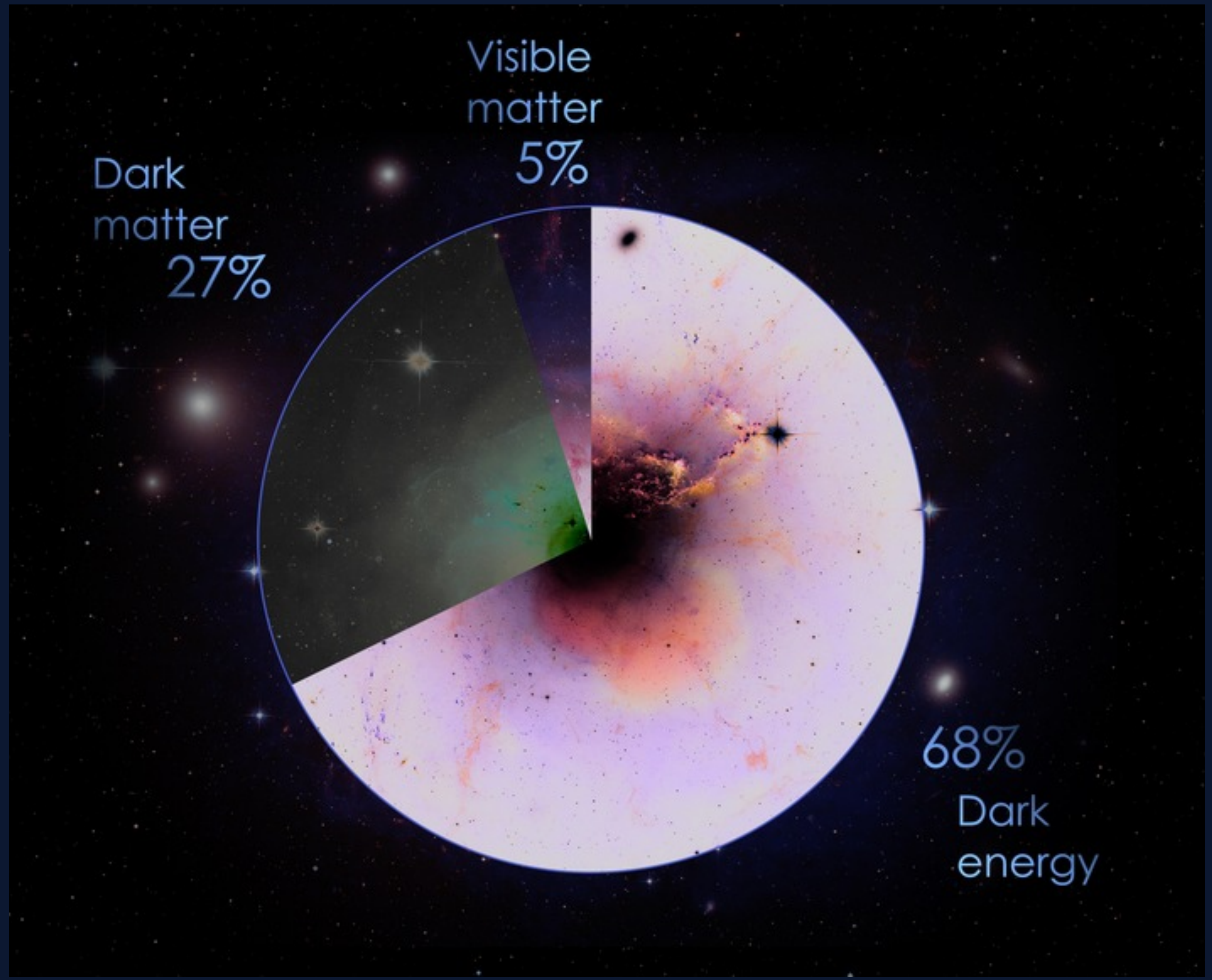


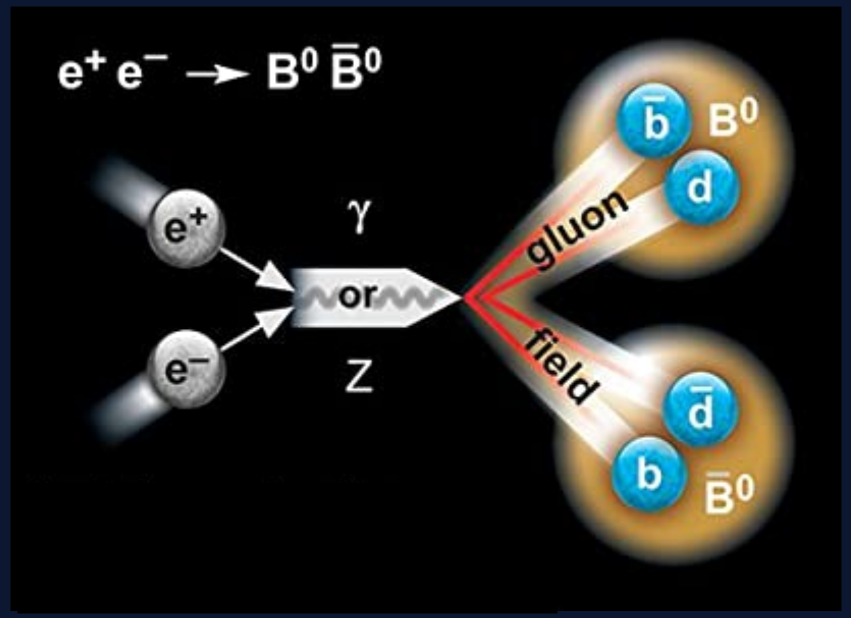


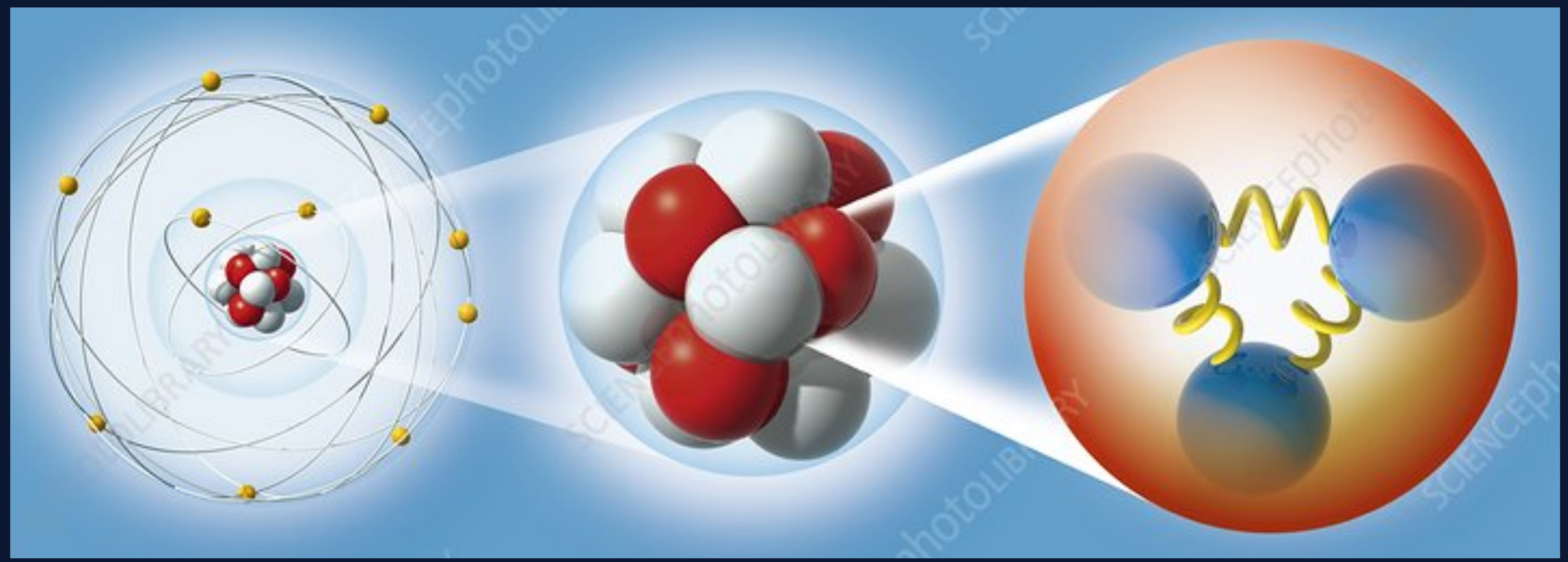


BIG QUESTIONS

- What is our universe made of?
- What are the smallest building blocks of nature?
- How do they interact with each other?
- Is our universe stable?





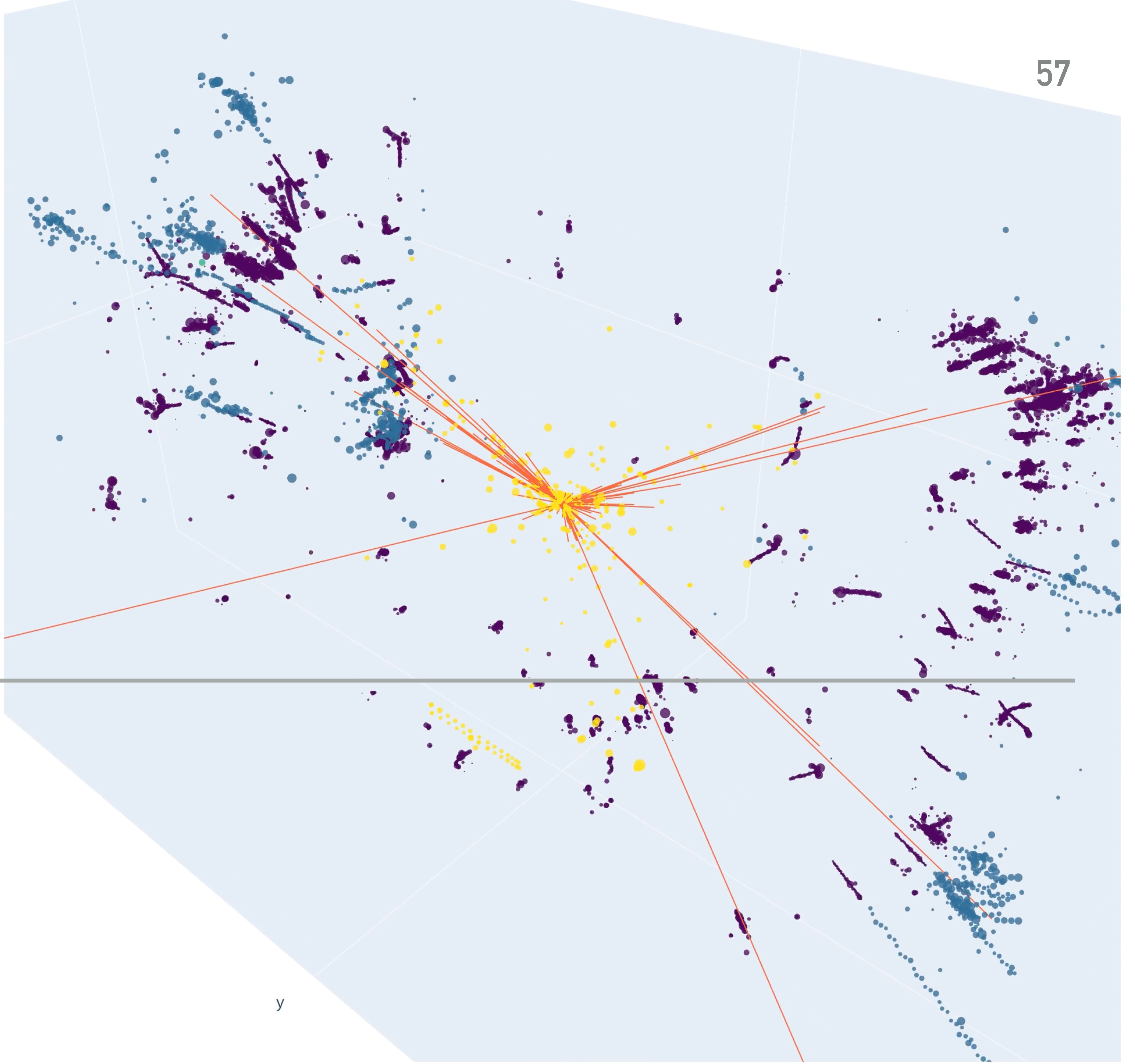


JAVIER DUARTE

ICML 2024

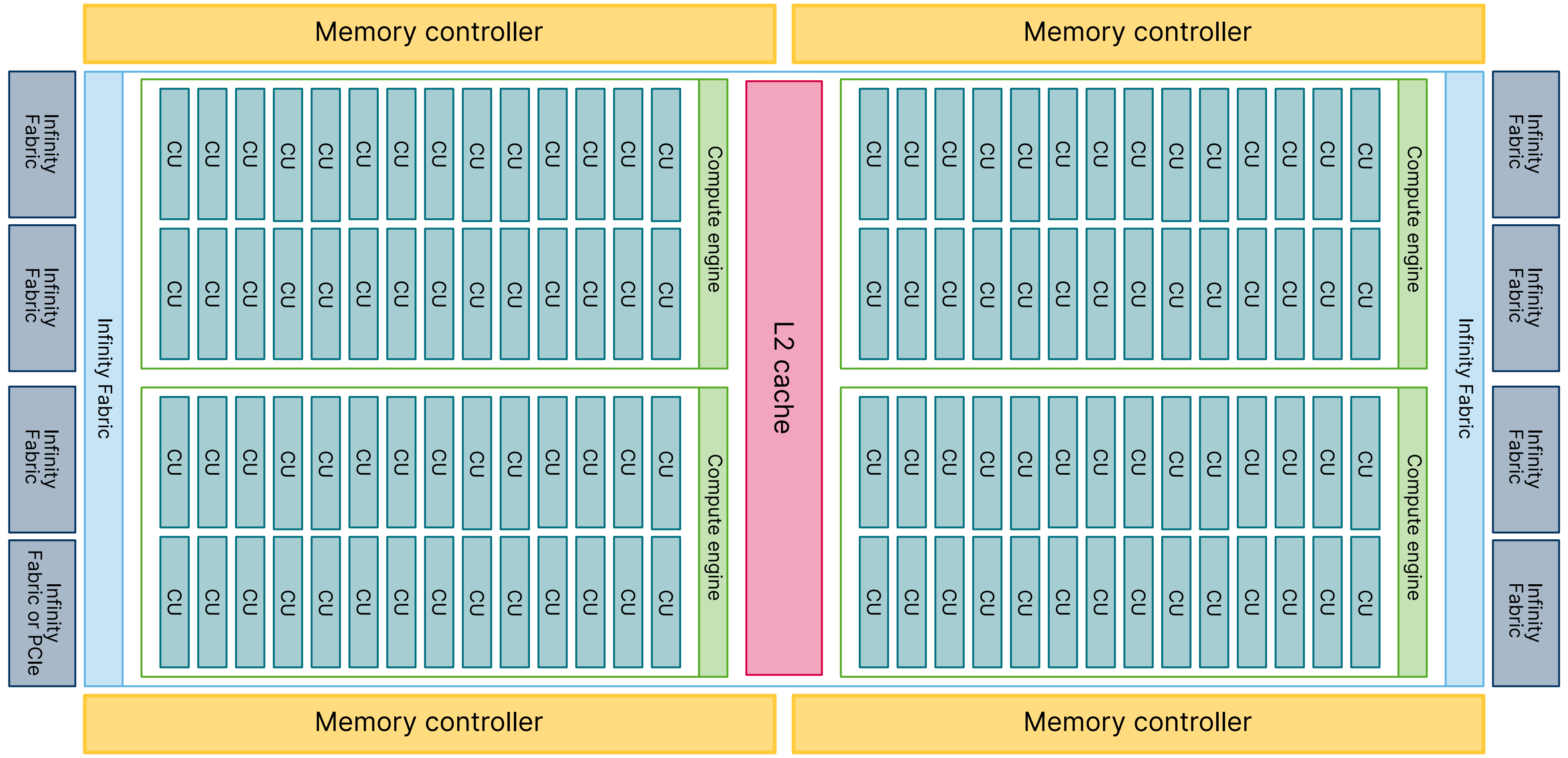
JULY 24, 2024

BACKUP

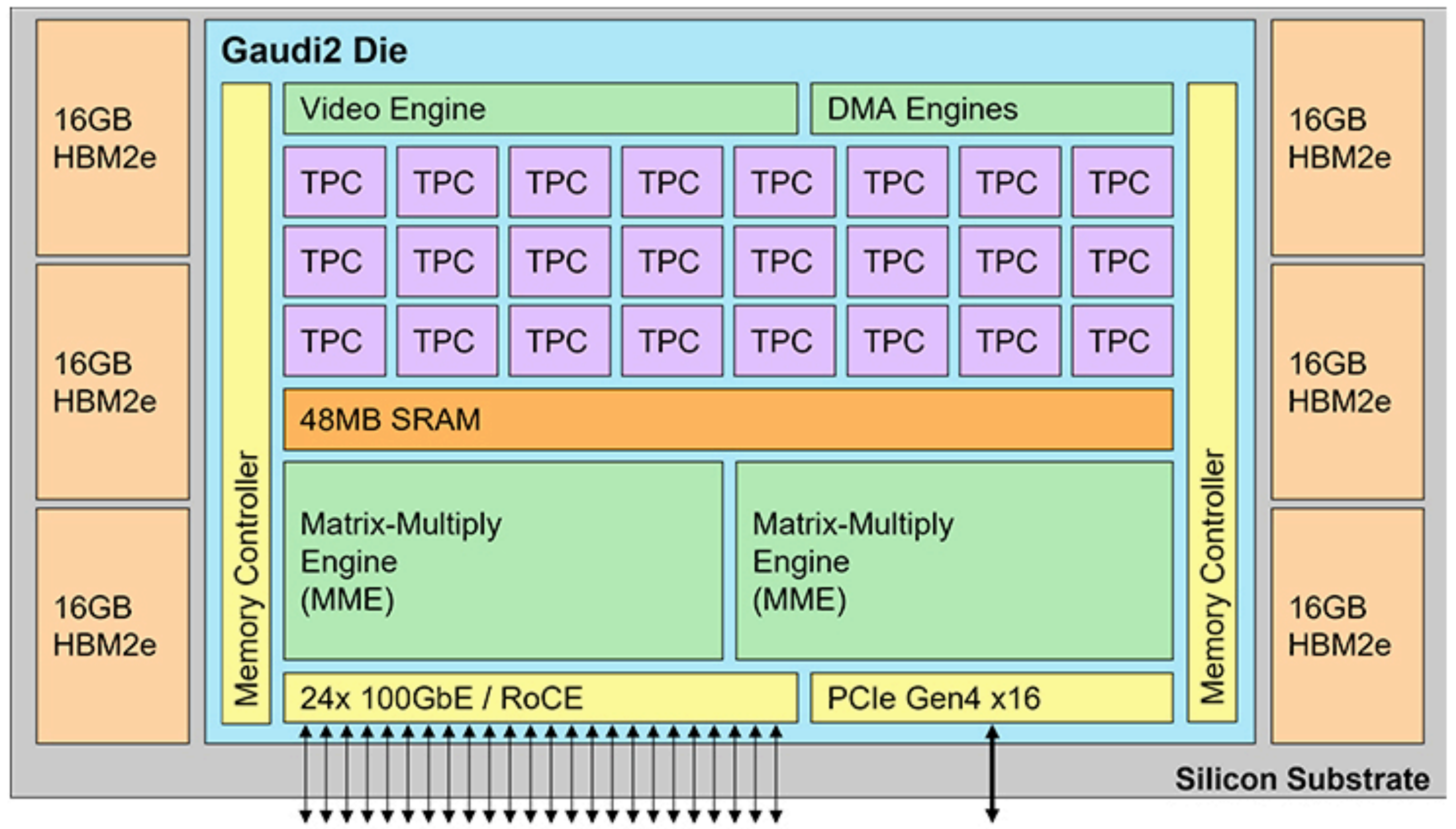


HPC AI CHIPS

The HPC AI chip landscape is diversifying



AMD MI250X GPU



Intel Gaudi2 deep learning processor

… we need flexible and portable codes to make use of these resources in the near future!

PORTABILITY

Portable on CPU, Nvidia & AMD GPU, Intel Habana Gaudi chips