

Artificial Intelligence and Machine Learning in Physics

Yeonju Go

Brookhaven National Laboratory / Stony Brook University

July 27, 2026

*BNL Physics Department Summer Lecture
Upton, New York, USA*



U.S. DEPARTMENT
of **ENERGY**

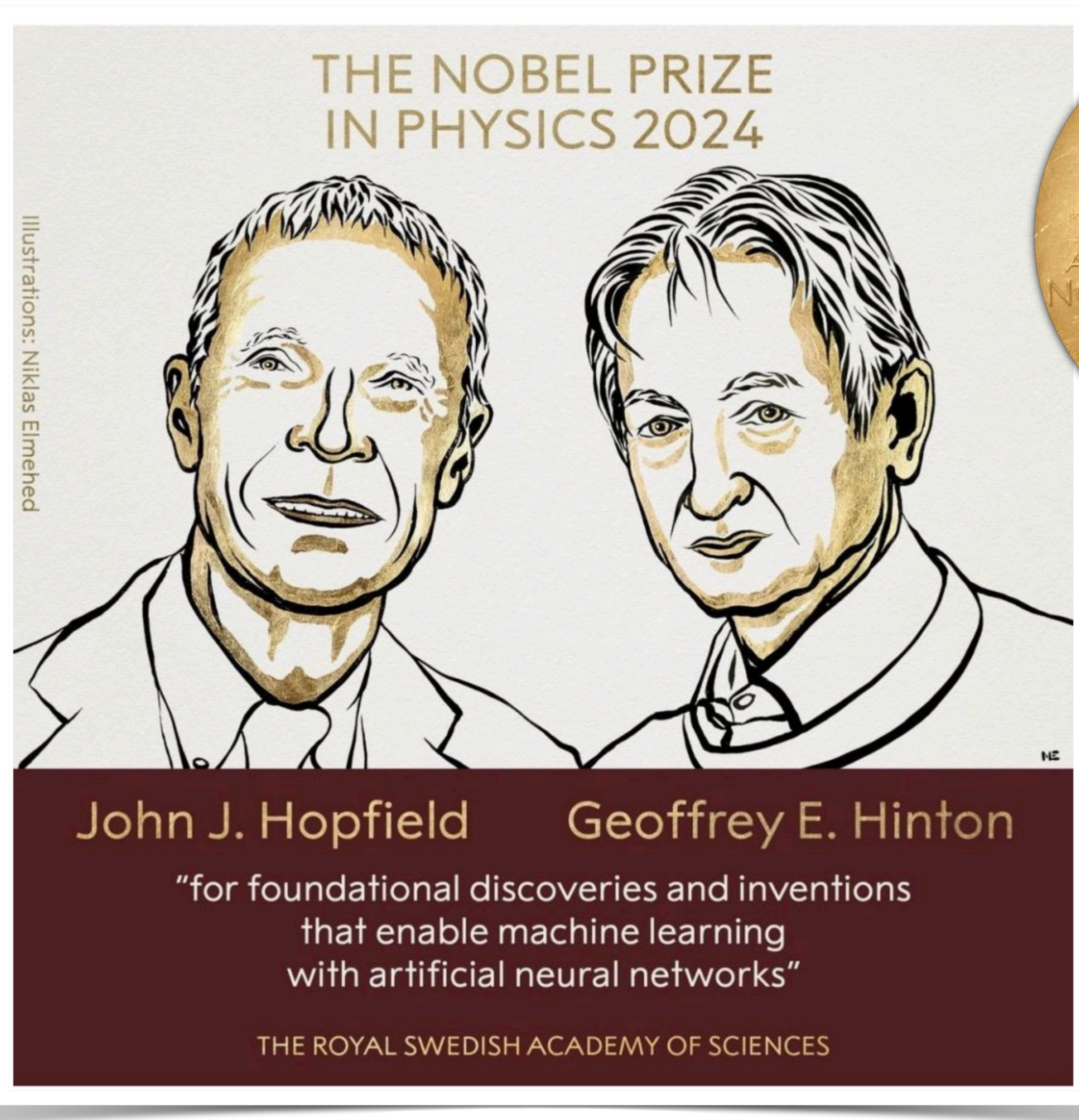


Brookhaven
National Laboratory



Stony Brook
University

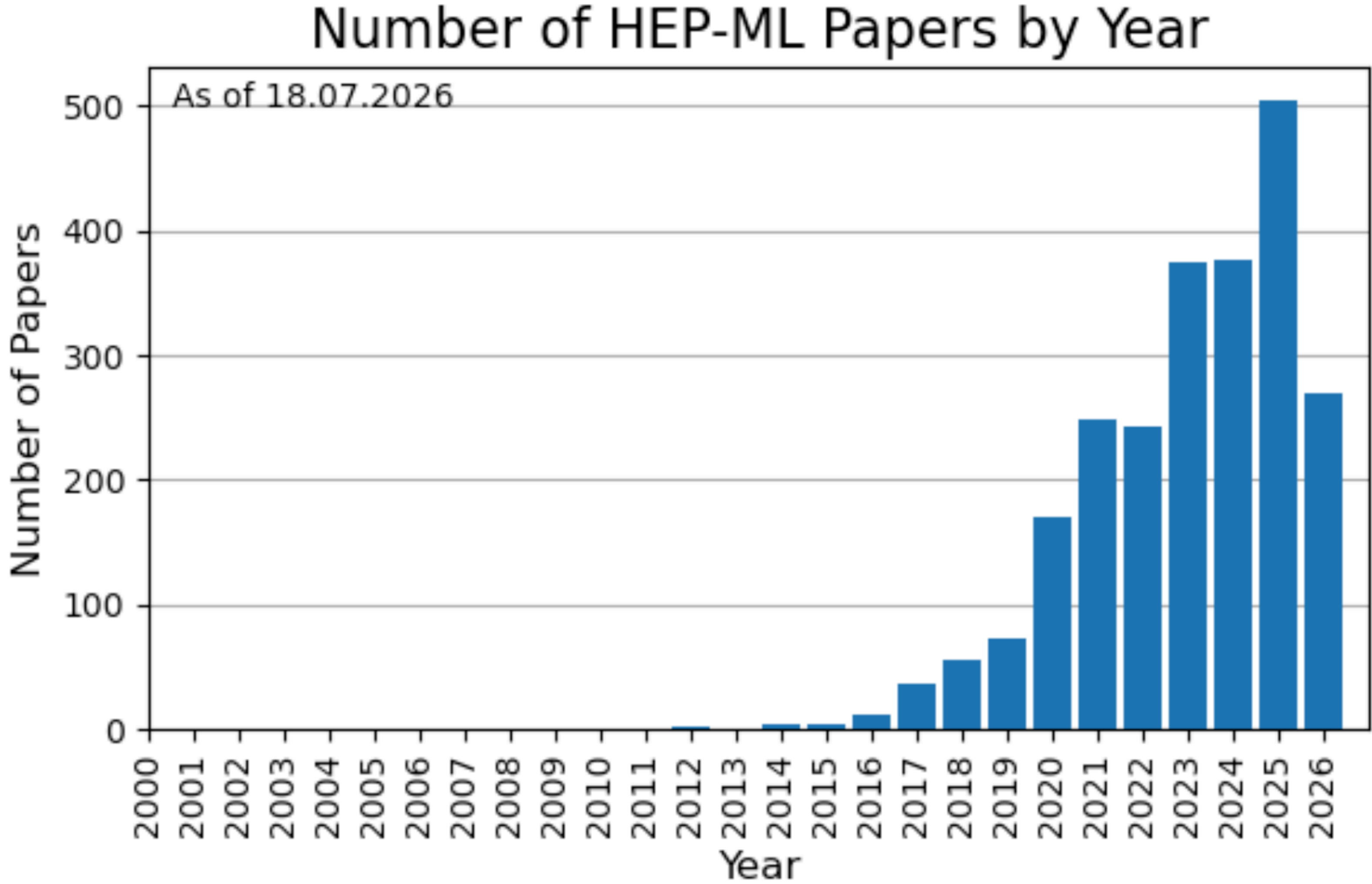
Artificial Intelligence and Machine Learning in Physics



2024
Nobel Prize in Physics

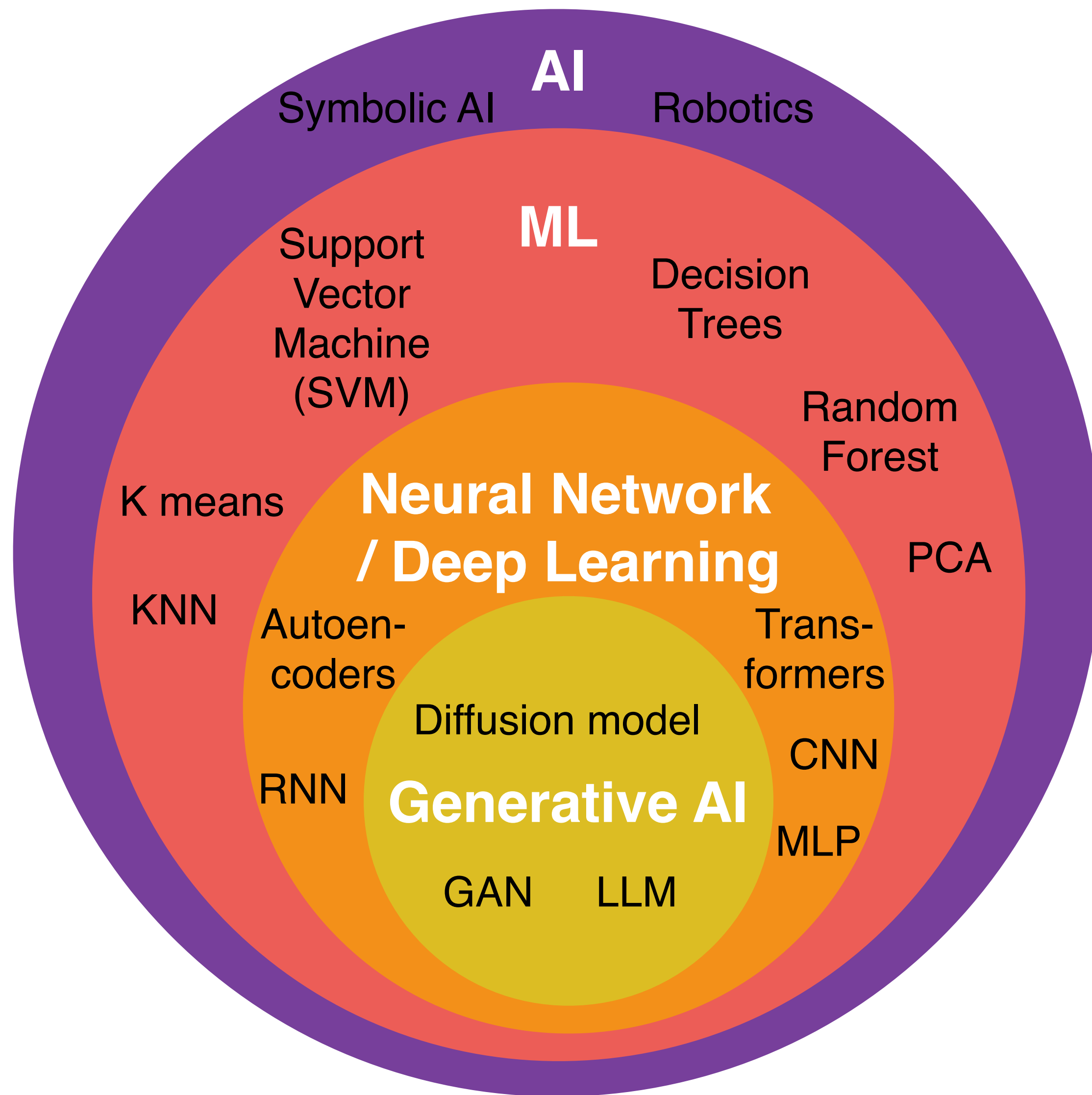
*"for foundational discoveries and
inventions that enable **machine learning**
with **artificial neural networks**"*

Rapid Growth of AI Publications in Physics



<https://iml-wg.github.io/HEPML-LivingReview/>

Artificial Intelligence and Machine Learning



Artificial Intelligence (AI)

- **mimic** cognitive functions associated with **human intelligence**
e.g. see, understand, respond, analyze data, make recommendations, etc

Machine Learning (ML)

- **extract knowledge from data** and learn from it **autonomously**
- **no explicit programming**

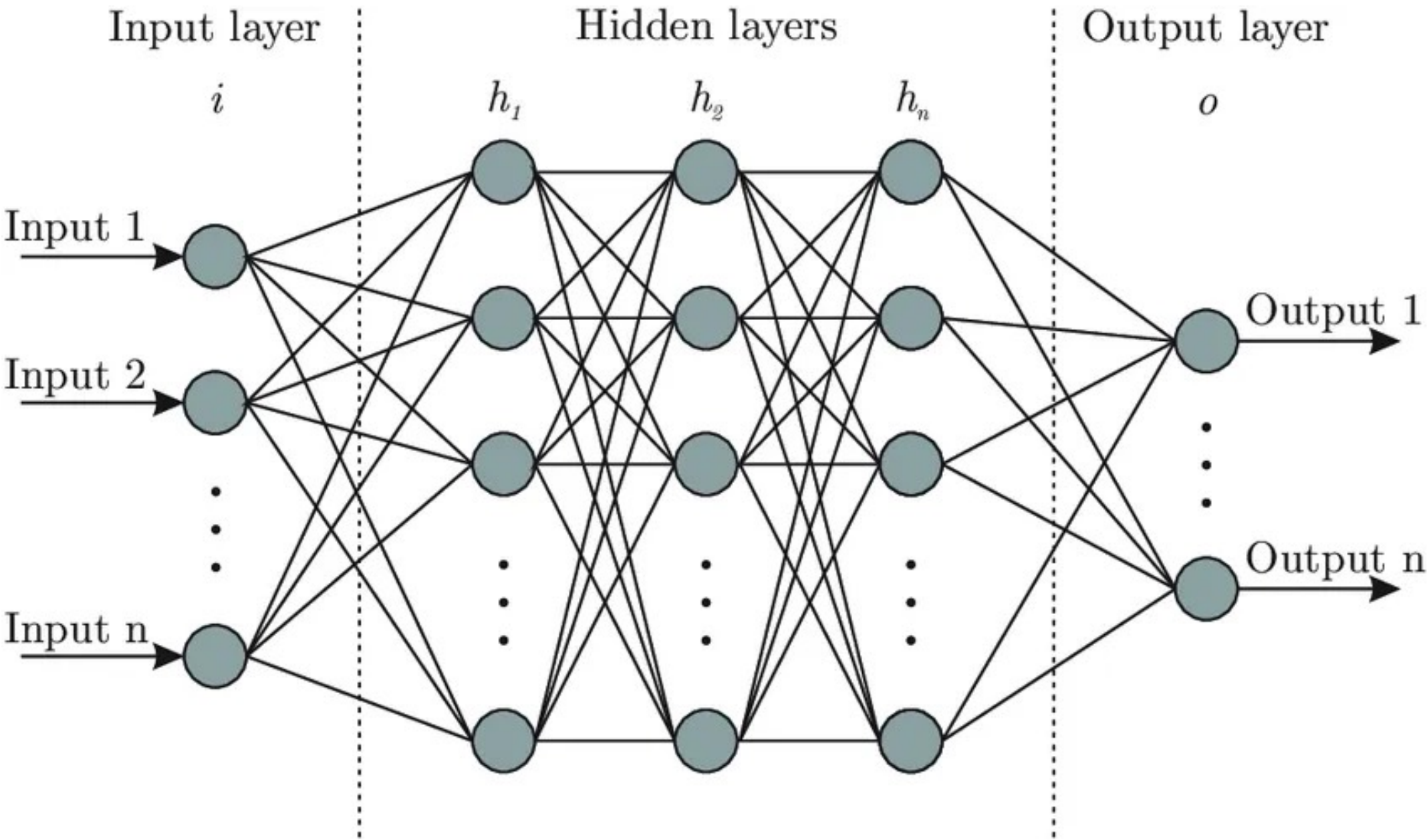
Neural Network / Deep Learning

- inspired by human brain that uses “**neurons**”
- learn based on deep neural network

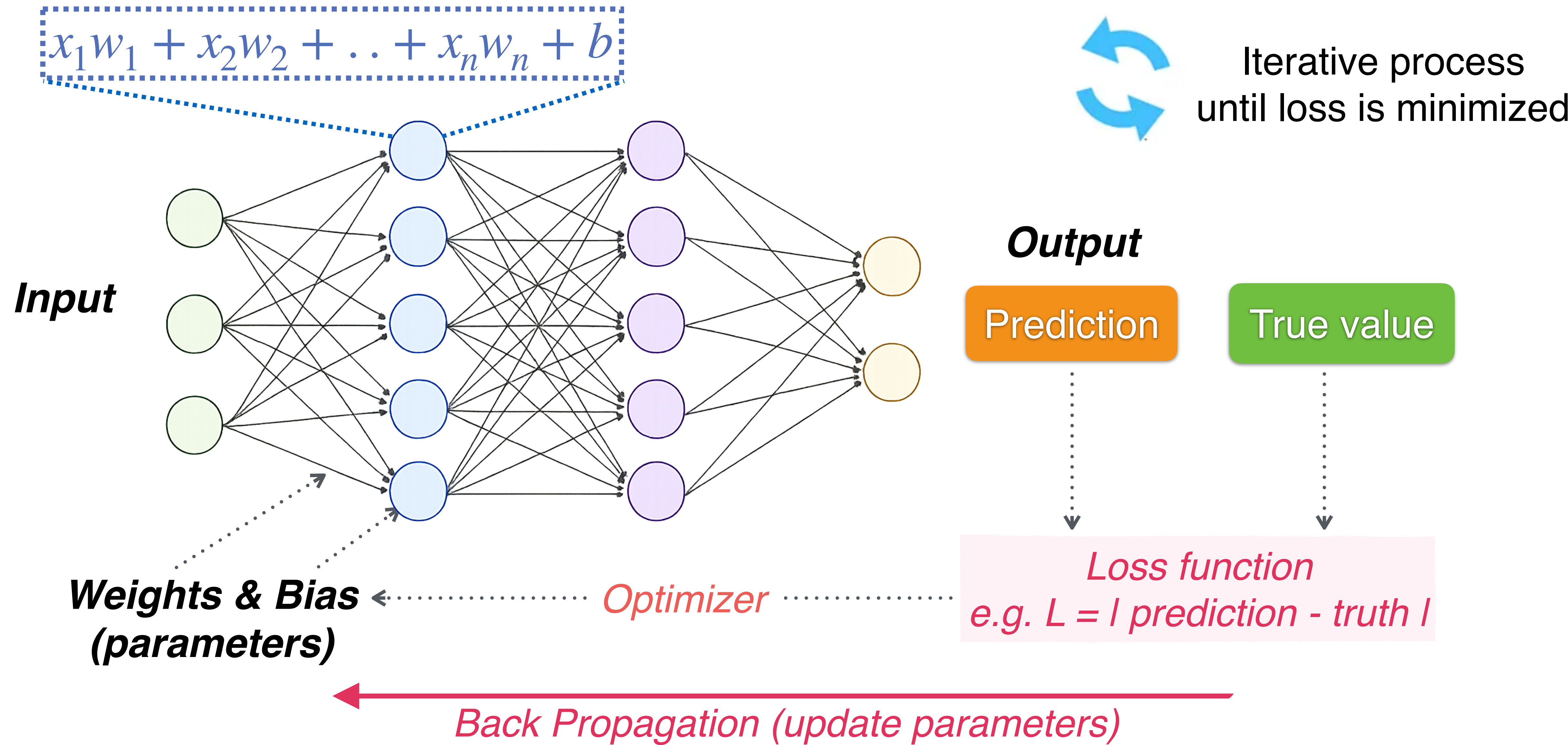
Generative AI

- **creates new content**, such as text, images, and audio, by learning patterns from existing data

Human Brain and Neural Network



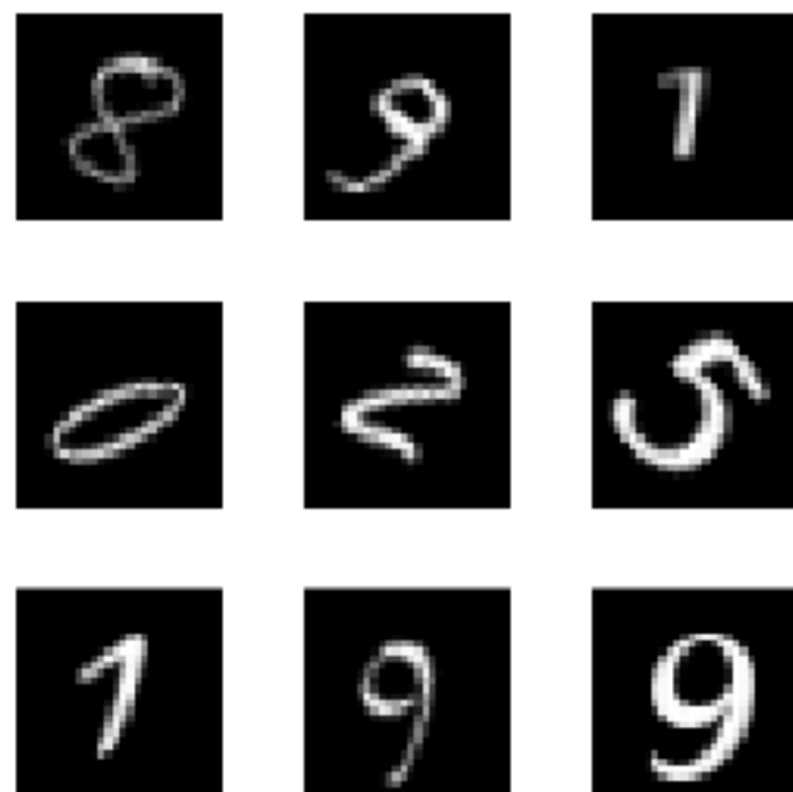
Neural Network



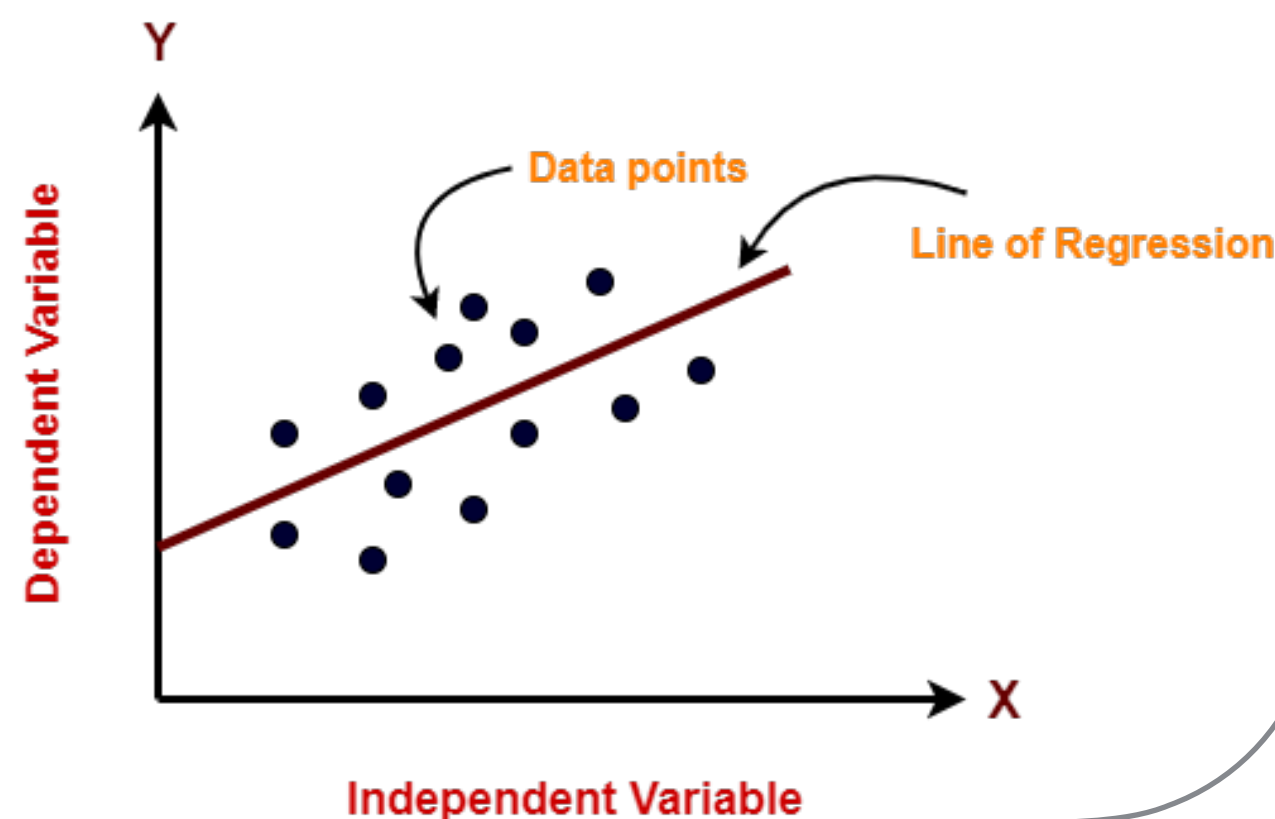
Types of Machine Learning

Supervised

- Use **labelled** data
- **Input-output values** are given by human
- Classification

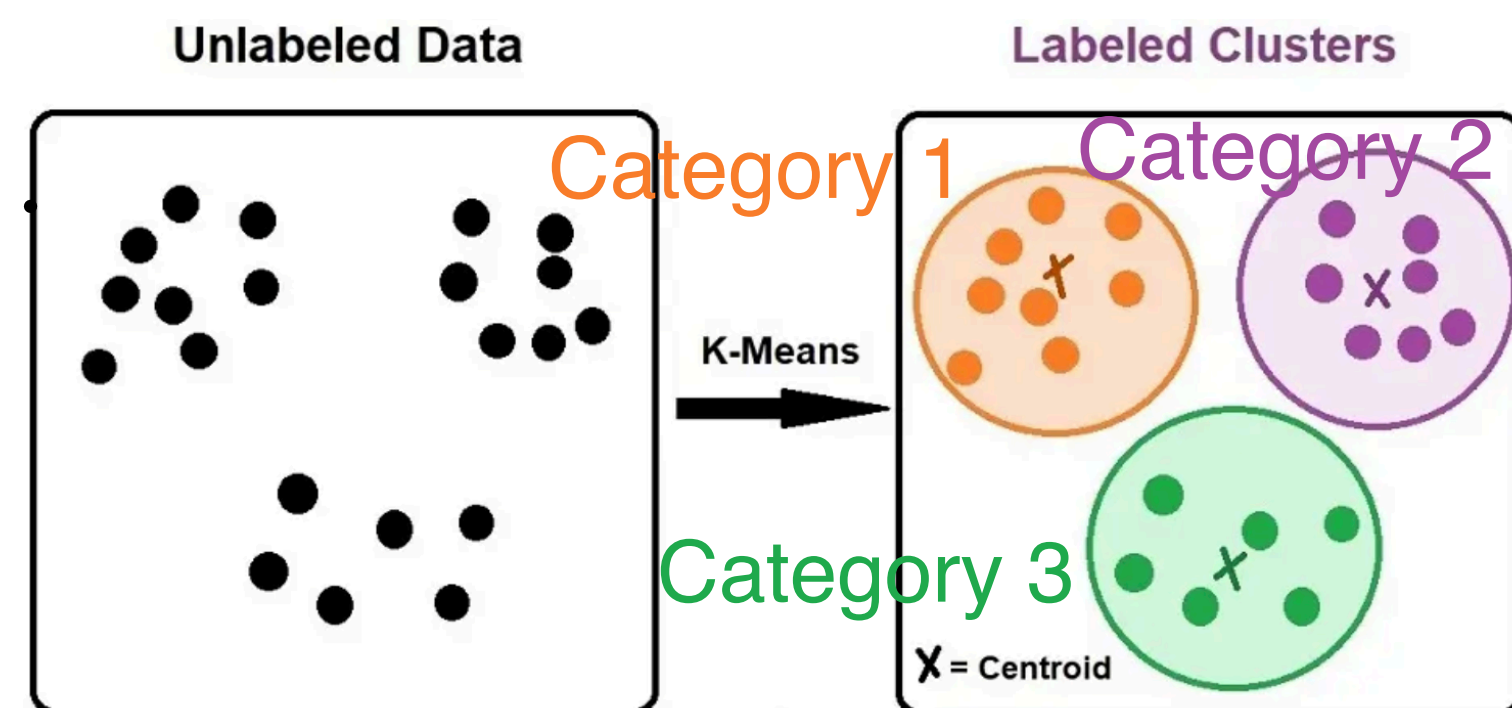


- Regression

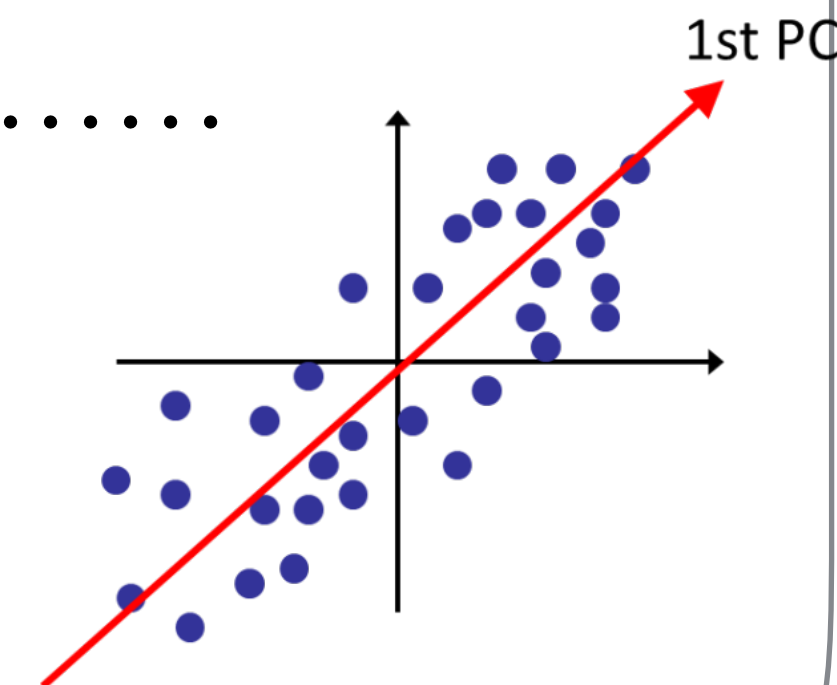


Unsupervised

- Use **unlabelled** data
- Find **hidden patterns** w/o human intervention
- Clustering...



- Dimension Reduction
- Association
- Anomaly detection



Different architectures are built from **neural network** components and differ in how they organize, connect, and process information

Convolutional Neural Network?

Graph Neural Network?

Recurrent Neural Network?

Transformer?

Auto-encoder?

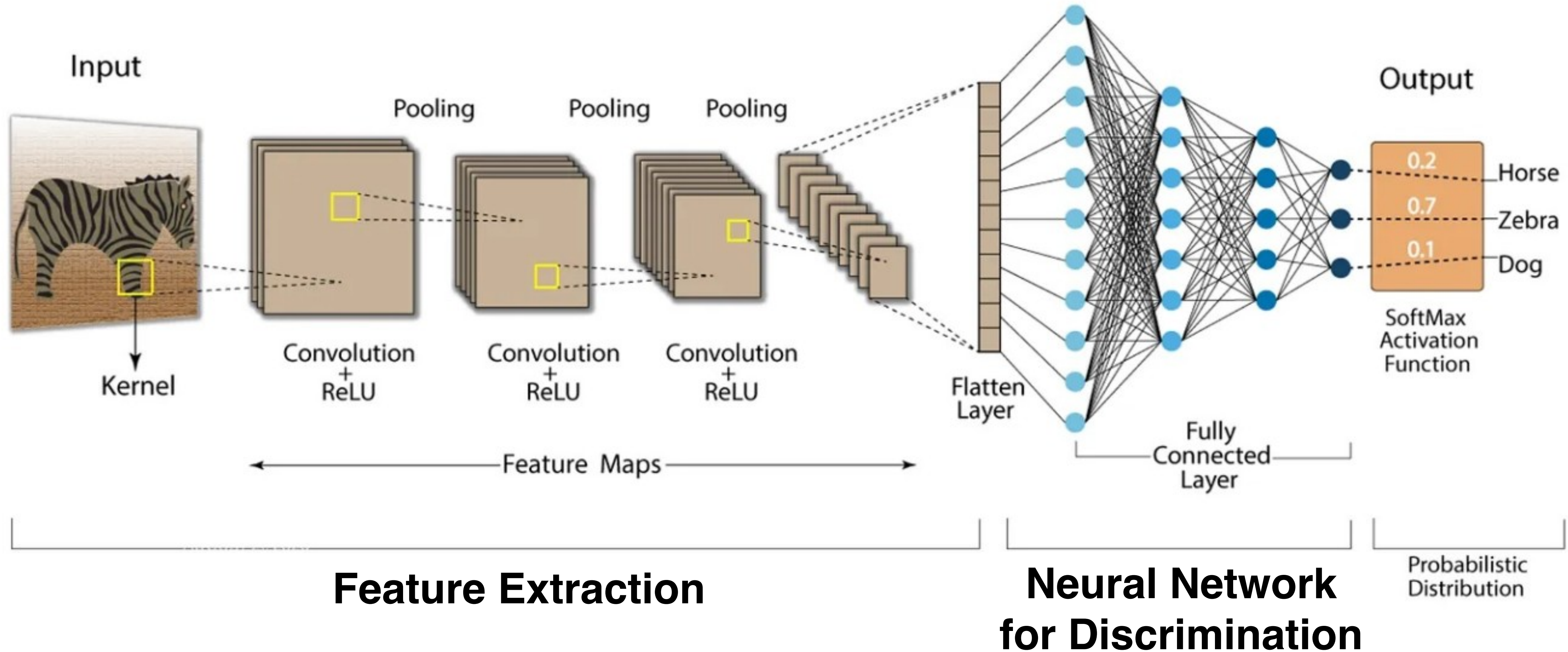
Diffusion Model?

Generative Adversarial Network?

...

Convolutional Neural Network (CNN)

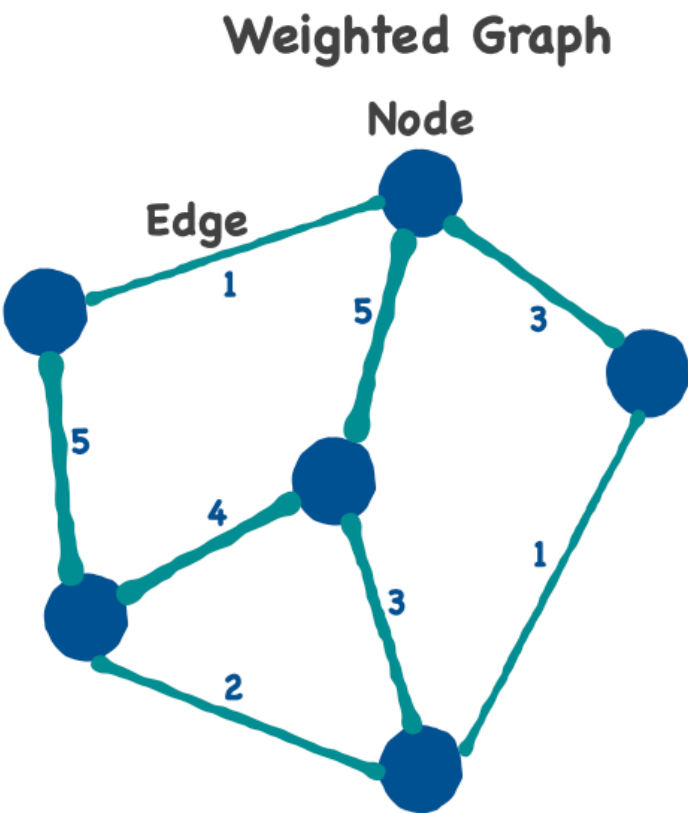
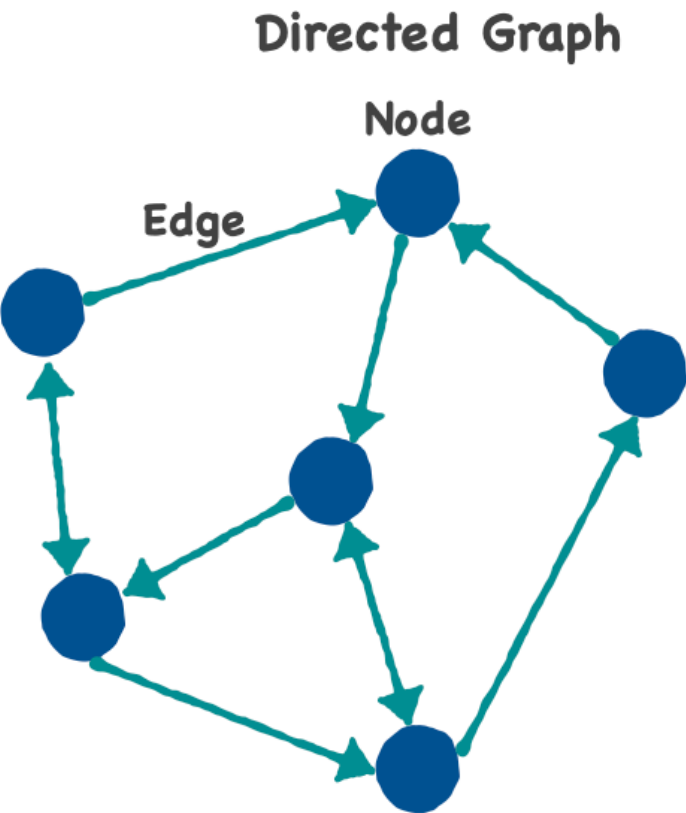
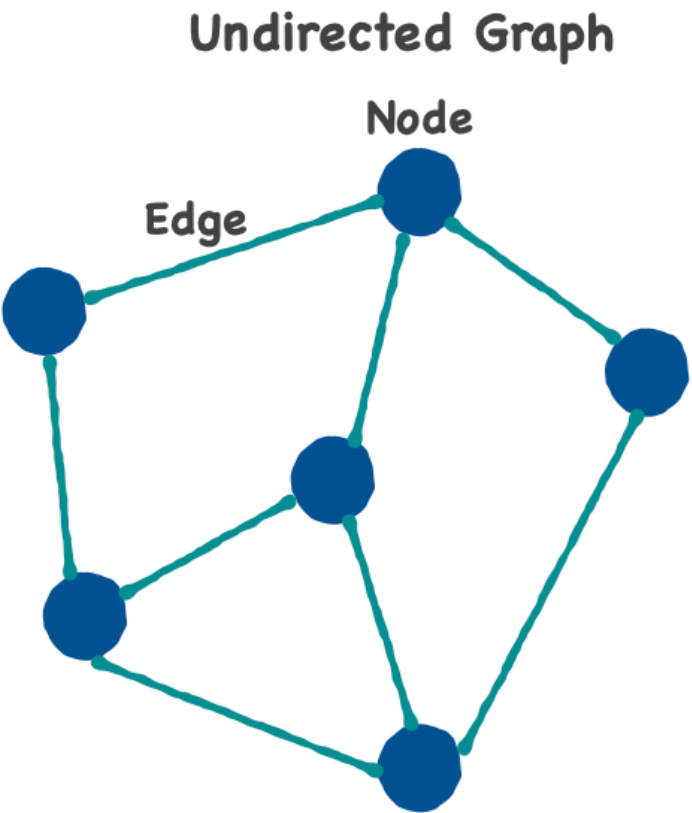
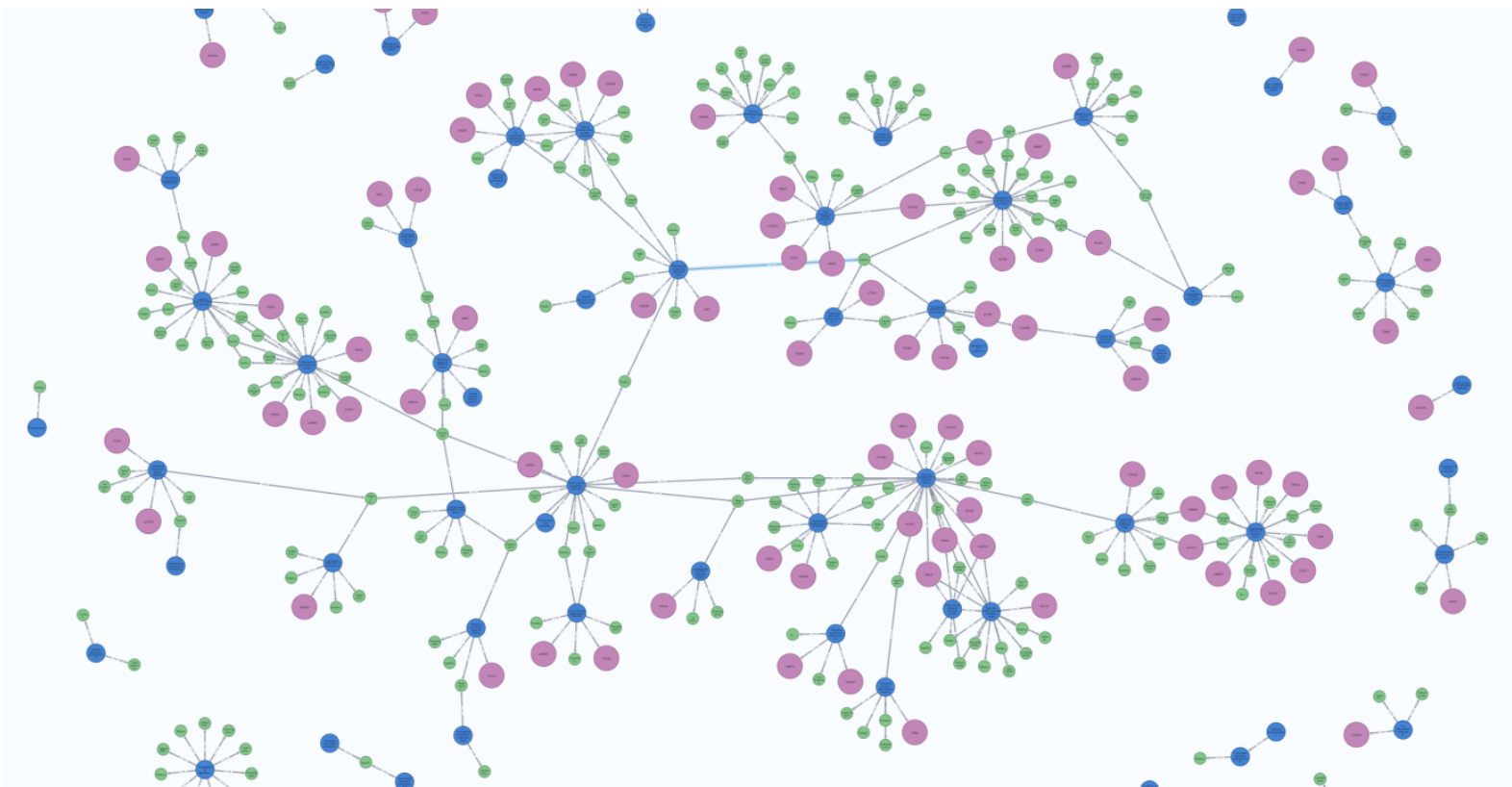
- Convolutional Neural Network (**CNN**) → good for **image** processing



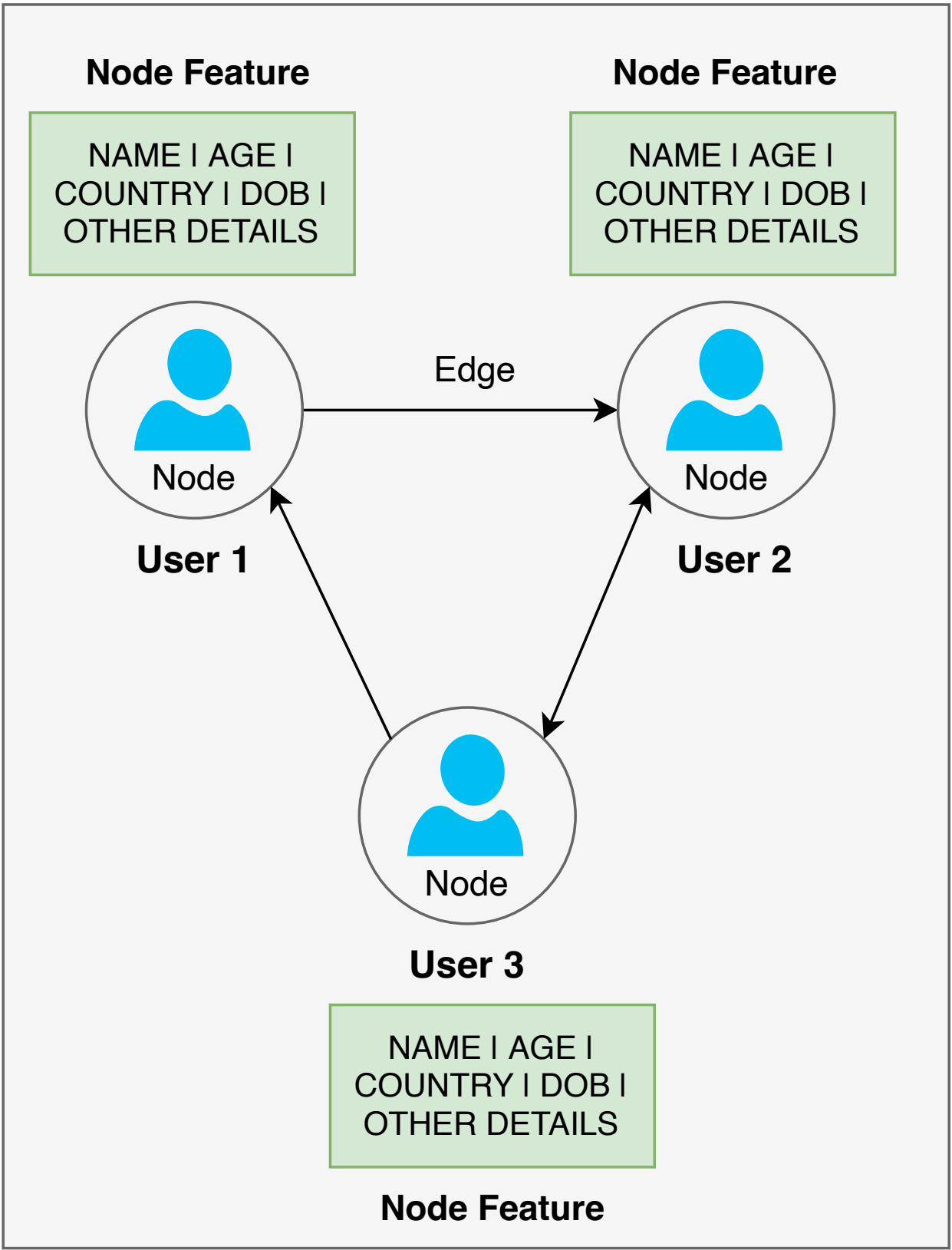
<https://developersbreach.com/convolution-neural-network-deep-learning/>

Graph Neural Network (GNN)

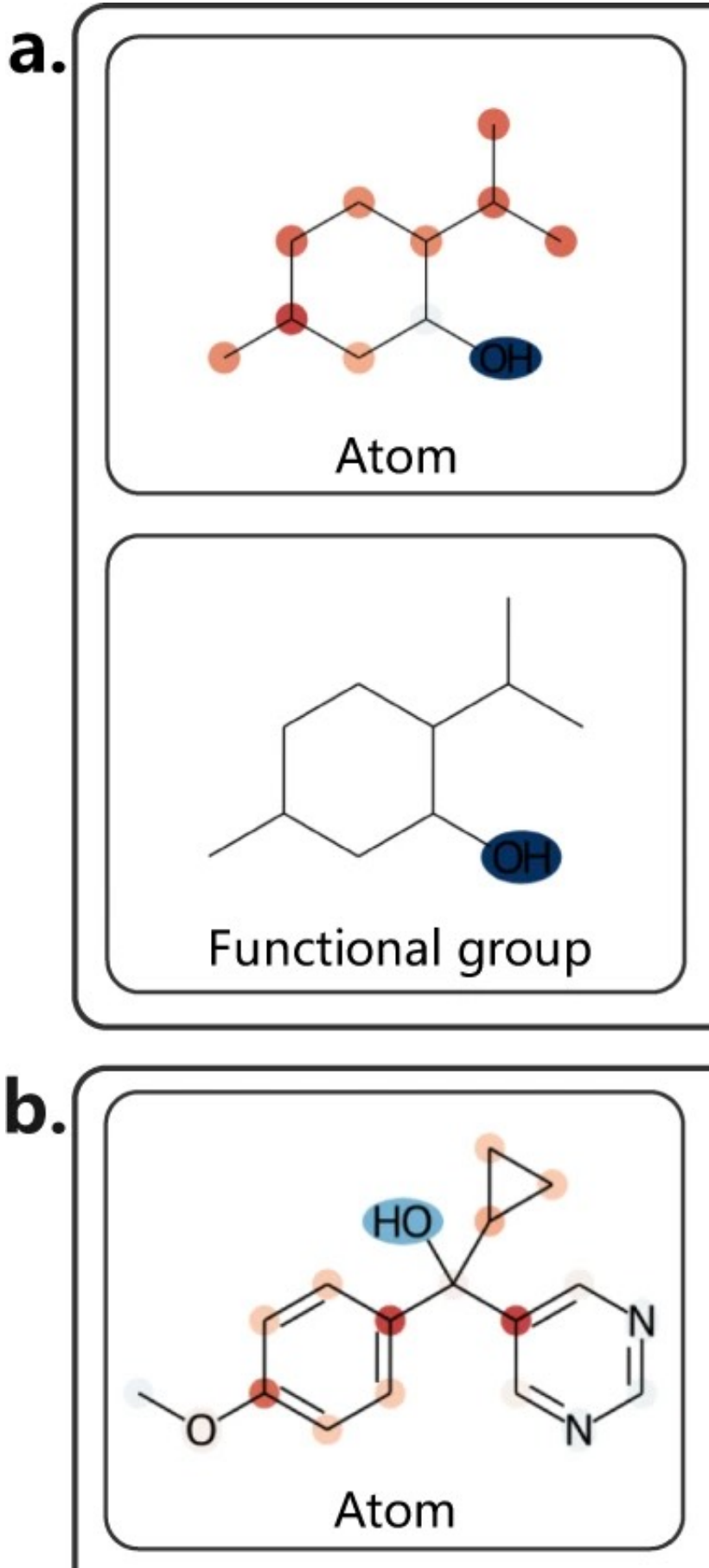
- Graph Neural Network (**GNN**) → good for **irregular, relational** data processing



<https://blogs.mathworks.com/deep-learning/2025/02/04/graph-neural-networks-in-matlab/>



<https://karthick.ai/blog/2024/Graph-Neural-Network/>



[Nature Communications 14, \(2023\) 2585](#)

Transformer

- Transformer uses attention mechanism, widely used for natural language processing (NLP)

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

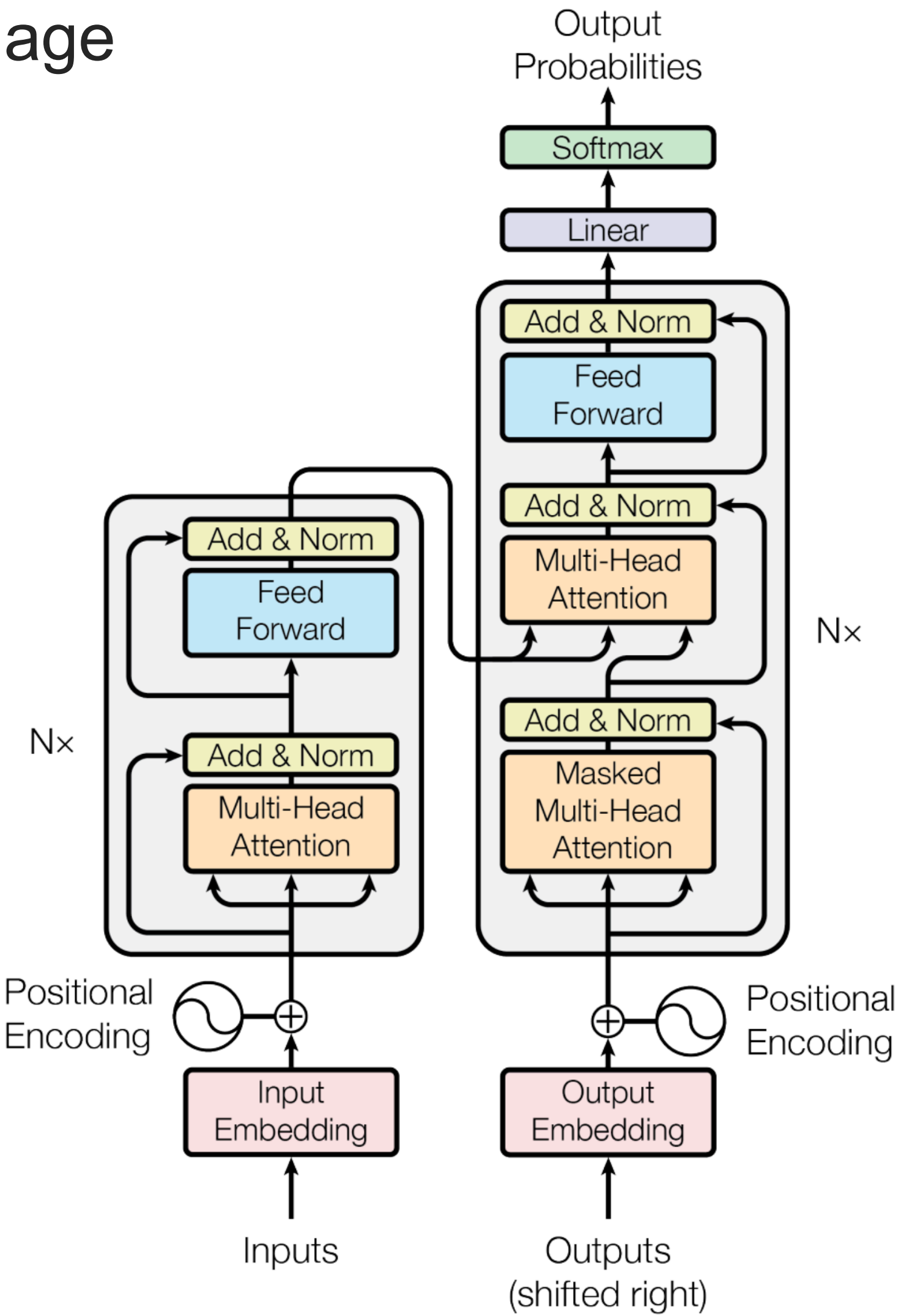
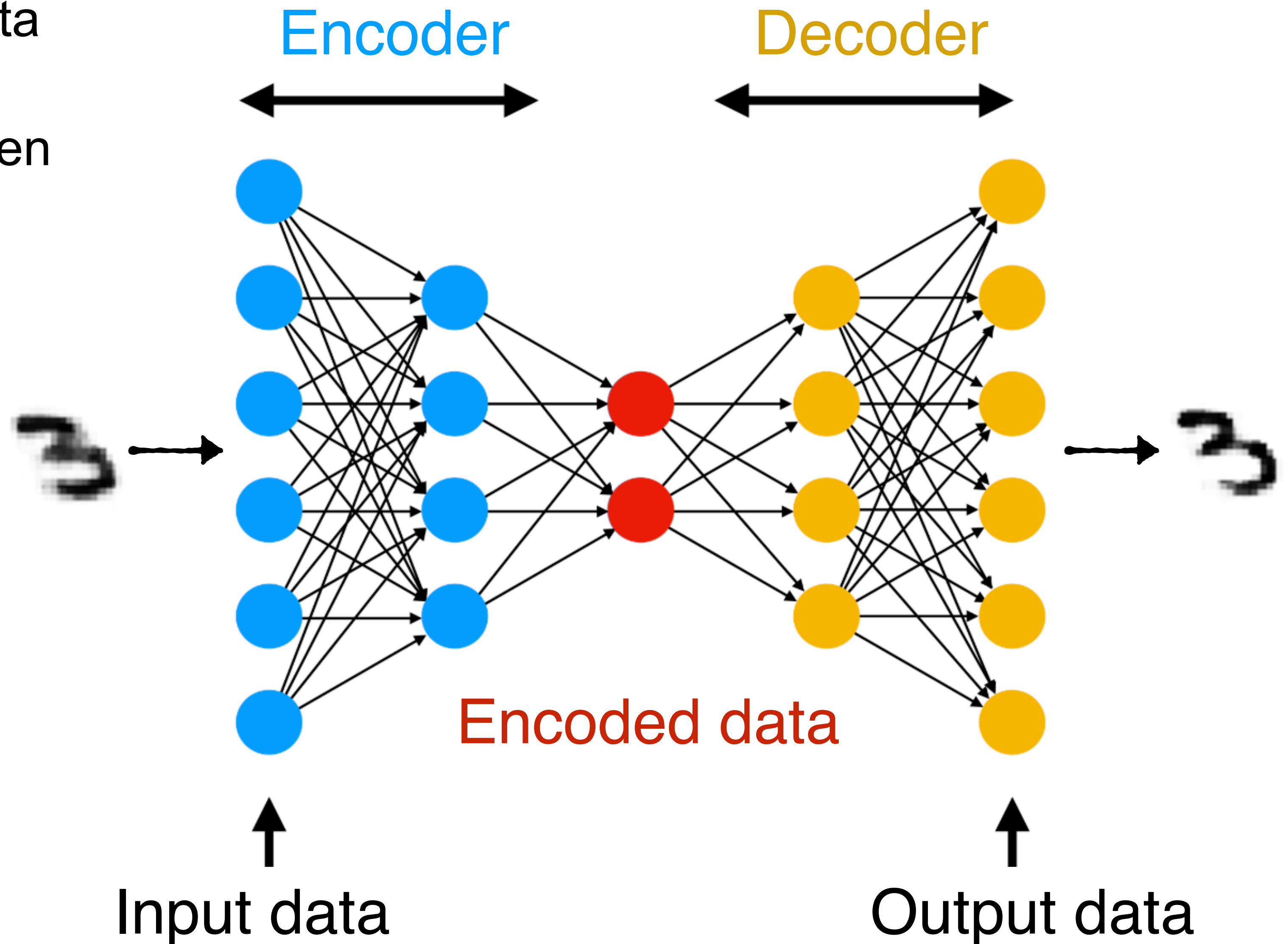


Figure 1: The Transformer - model architecture.

AutoEncoder

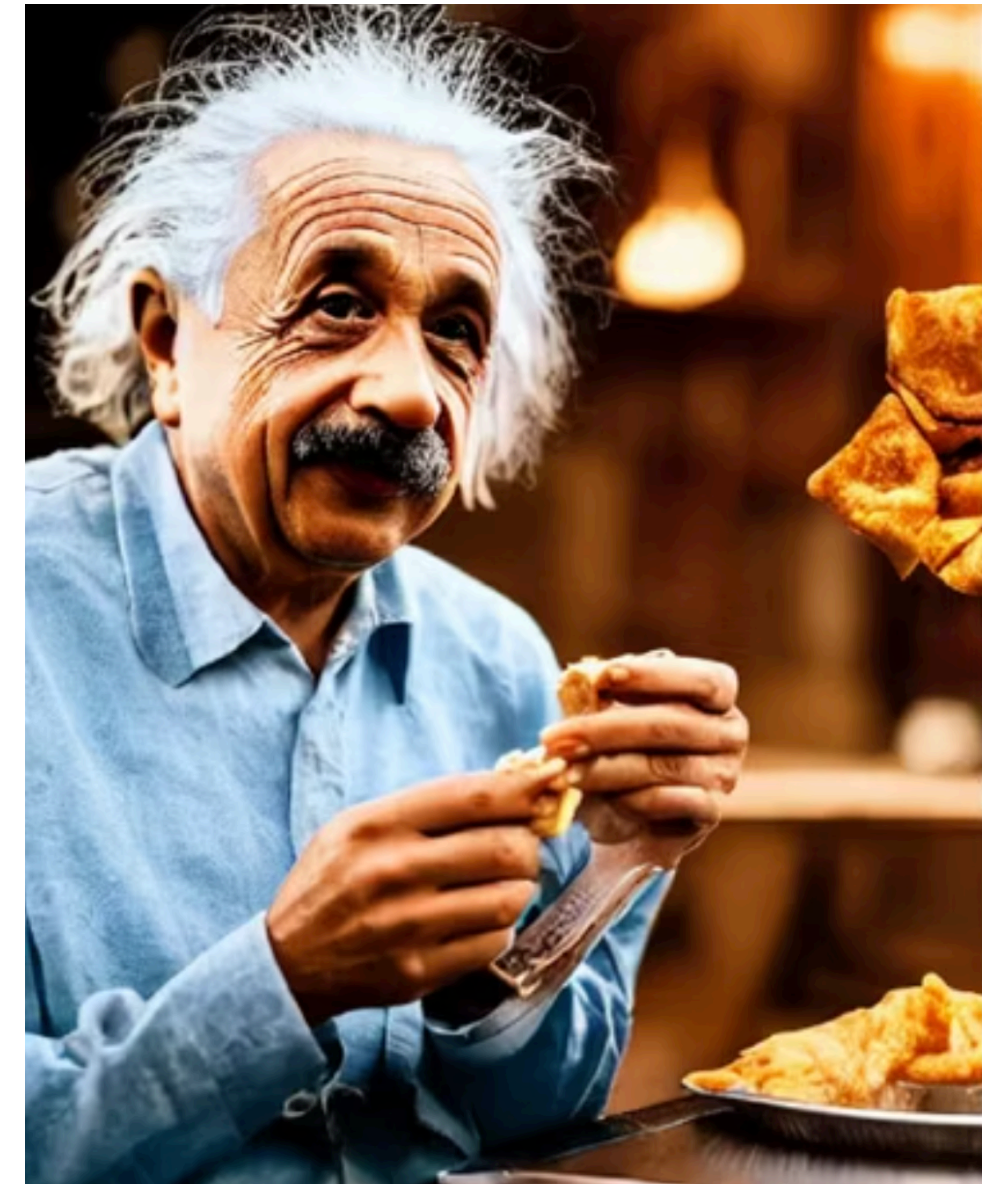
- **Autoencoder**: compress input data into a lower-dimensional representation (**encoding**) and then reconstruct it (**decoding**)
- Application:
 - *image denoising*
 - *data compression*
 - *feature learning*
 - *anomaly detection*
 - *dimensionality reduction*
 - ...



Generative AI

- *Generates syntactic data e.g. text, image, audio, video, ...*
- *Generally more advanced and complex AI*
 - ➔ combines representation learning (understanding) with generation (creation)
- *Models*
 - ➔ Generative Adversarial Network (GAN)
 - ➔ Diffusion models
 - ➔ Variational Autoencoder (VAE)
 - ➔ Transformer-based models, e.g. Large Language Model
 - ➔ ...

Generated by Dall-E 2

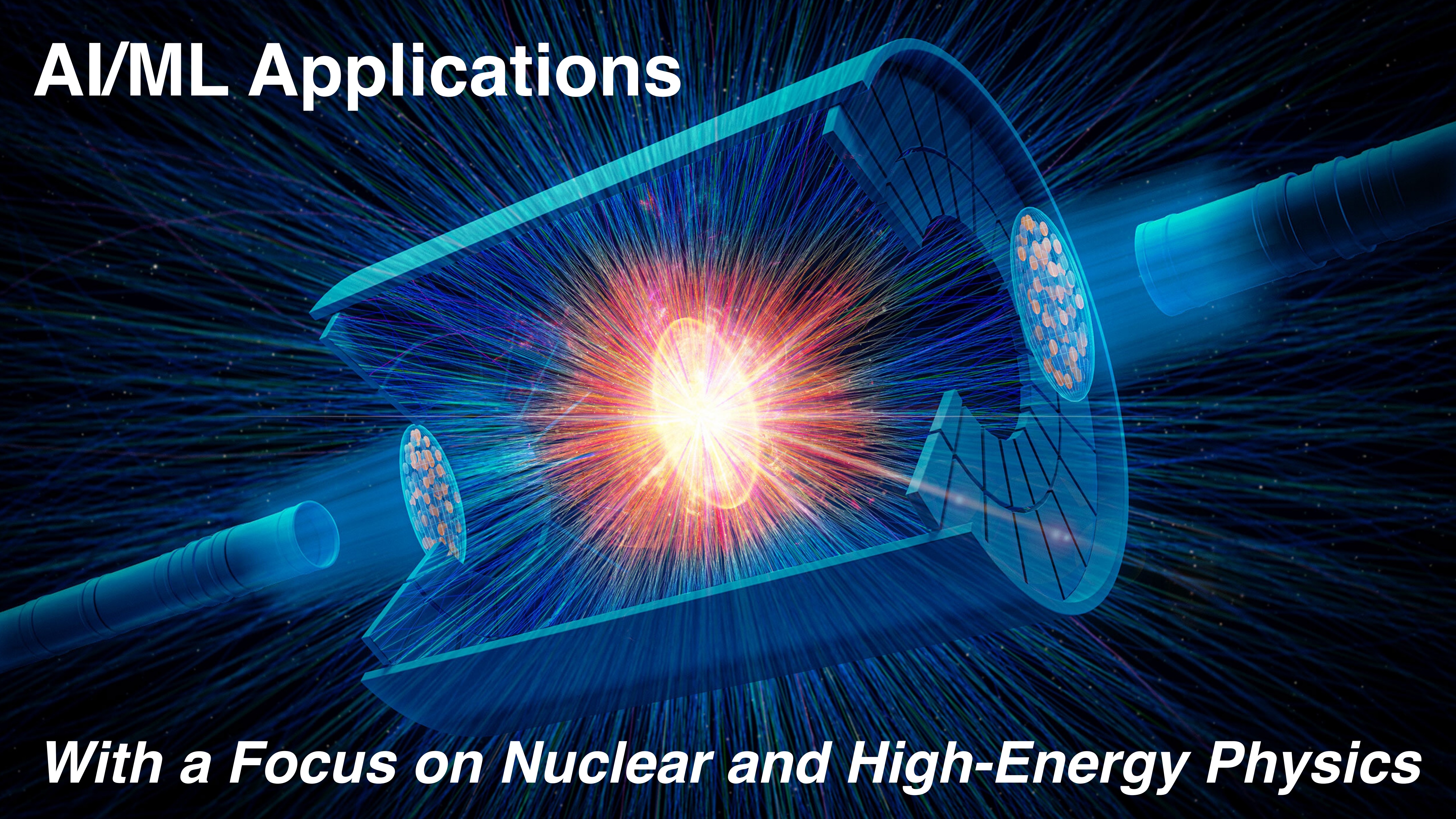




Which one is AI?



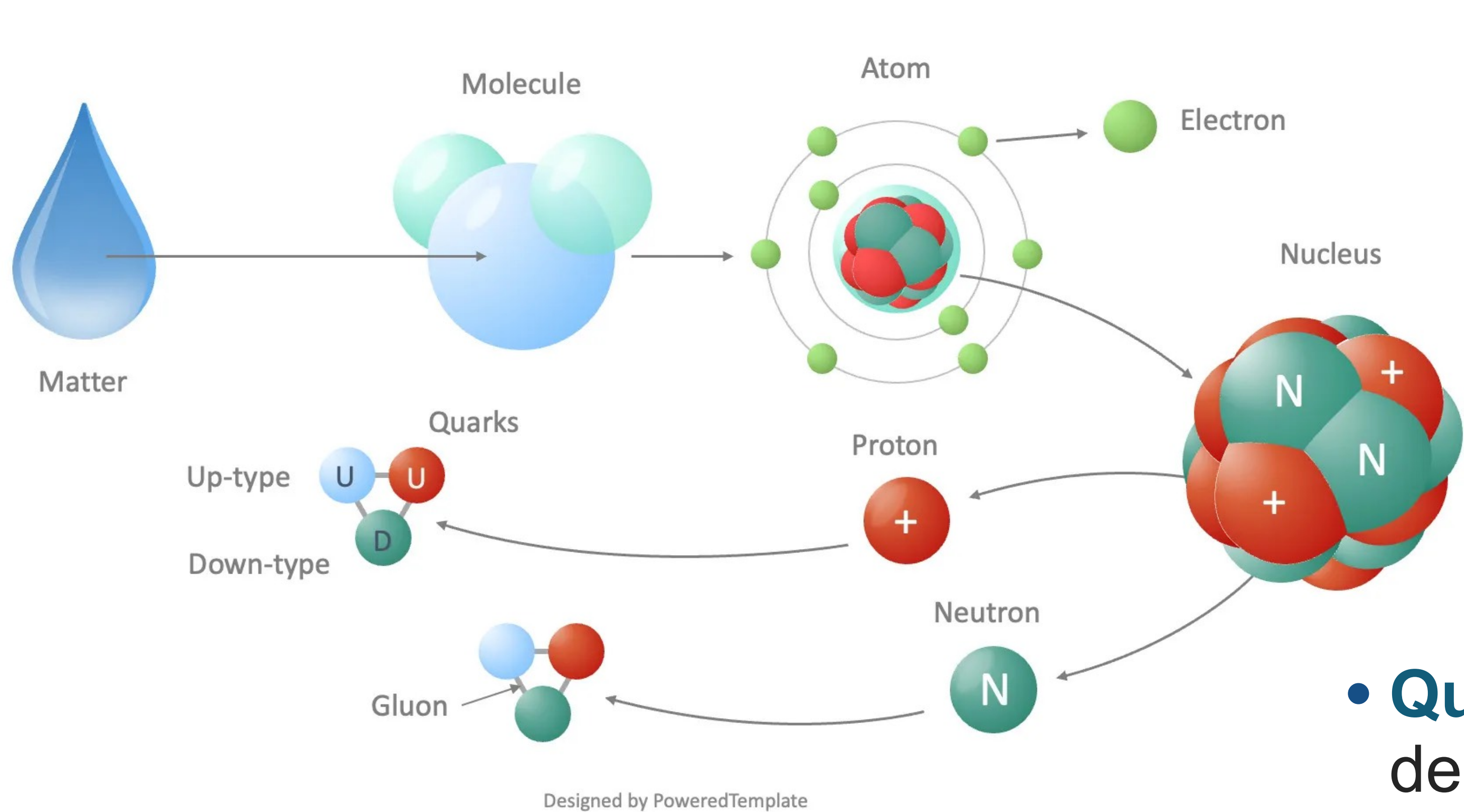
AI/ML Applications



With a Focus on Nuclear and High-Energy Physics

Quantum Chromodynamics

Matter from Molecule to Quark

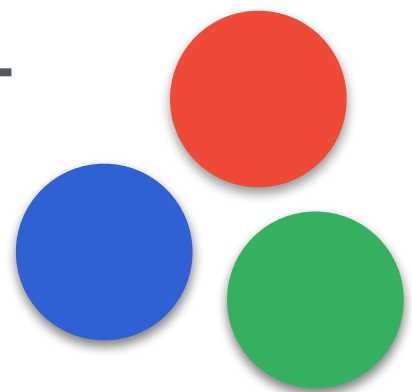


building blocks of matter

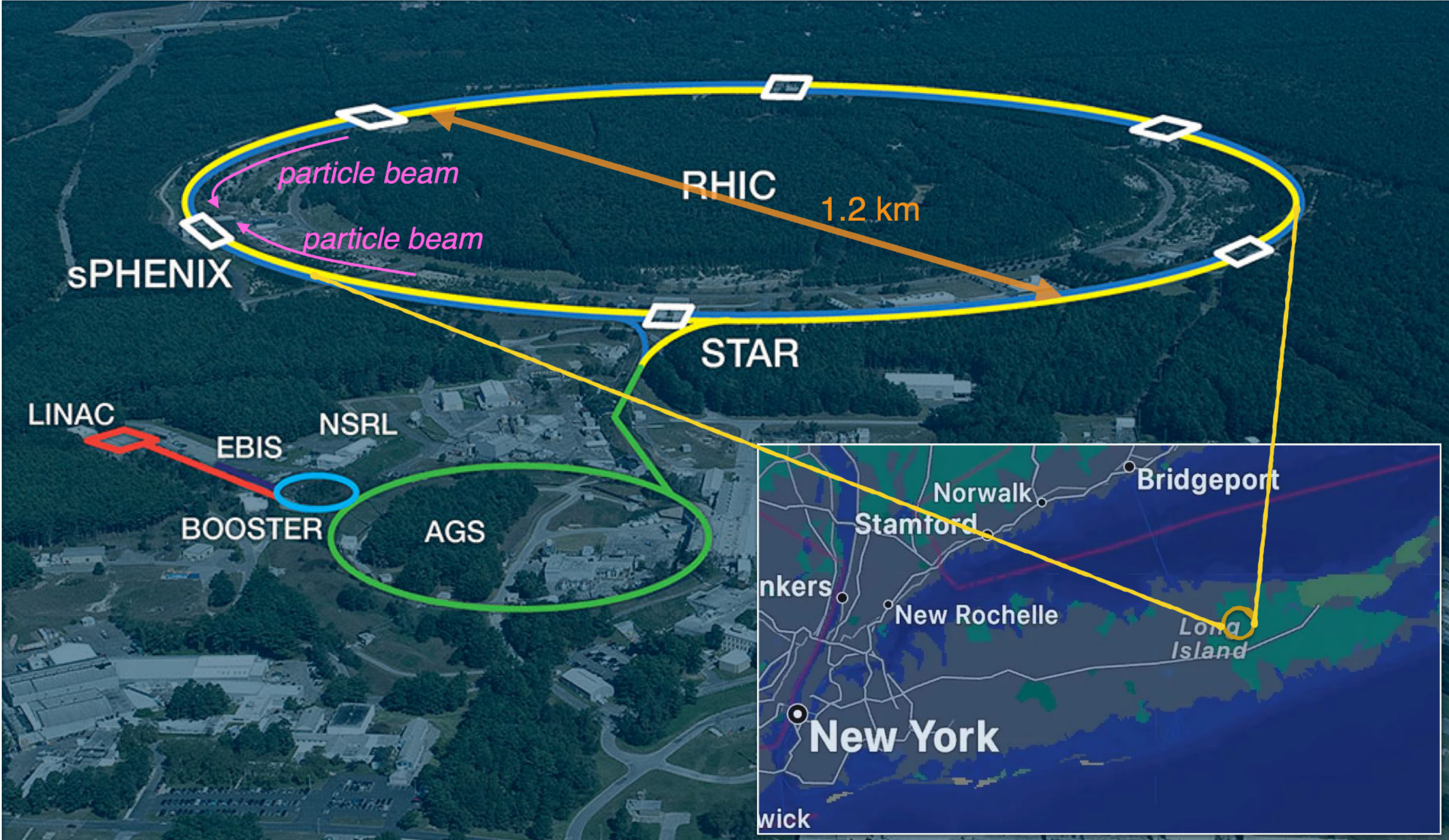
QUARKS	mass charge spin	$\approx 2.2 \text{ MeV}/c^2$ $\frac{2}{3}$ $\frac{1}{2}$	$\approx 1.28 \text{ GeV}/c^2$ $\frac{2}{3}$ $\frac{1}{2}$	$\approx 173.1 \text{ GeV}/c^2$ $\frac{2}{3}$ $\frac{1}{2}$	0 0 1
		u up	c charm	t top	g gluon
		$\approx 4.7 \text{ MeV}/c^2$ $-\frac{1}{3}$ $\frac{1}{2}$	$\approx 96 \text{ MeV}/c^2$ $-\frac{1}{3}$ $\frac{1}{2}$	$\approx 4.18 \text{ GeV}/c^2$ $-\frac{1}{3}$ $\frac{1}{2}$	
		d down	s strange	b bottom	

- Quantum Chromodynamics (QCD) describes the strong interaction between **quarks** and **gluons**

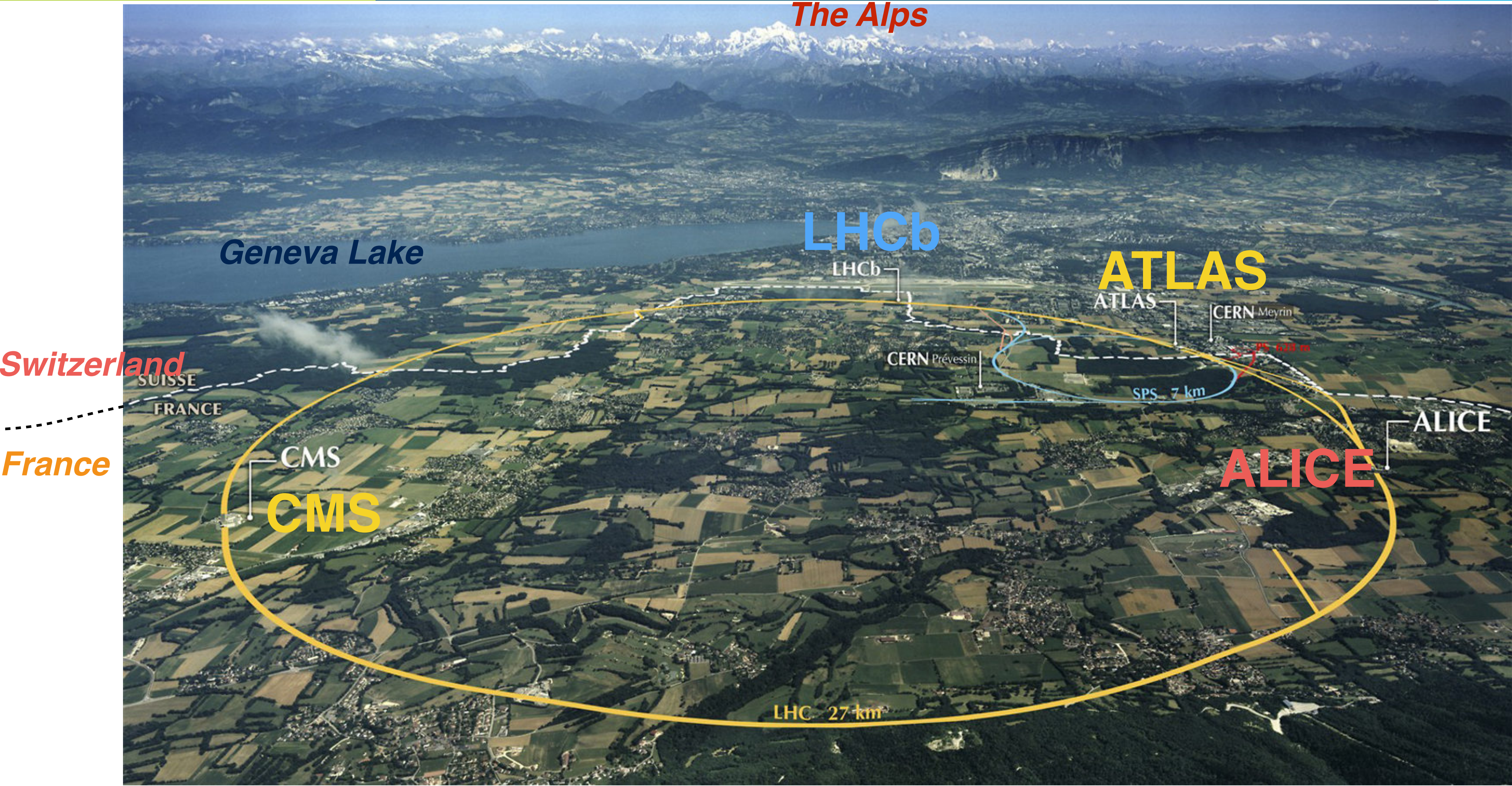
carry color charges



Relativistic Heavy Ion Collider (RHIC)

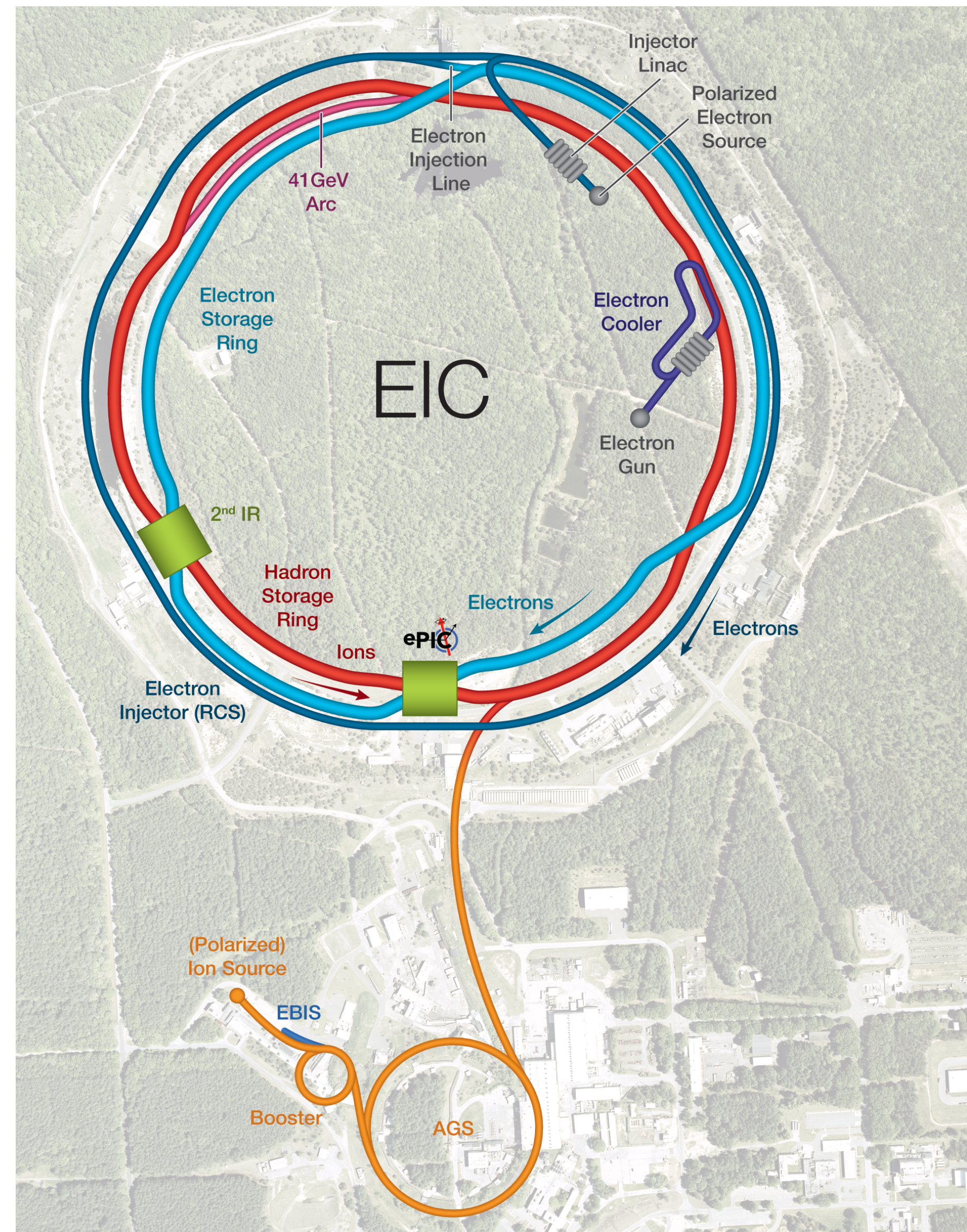


Large Hadron Collider (LHC)



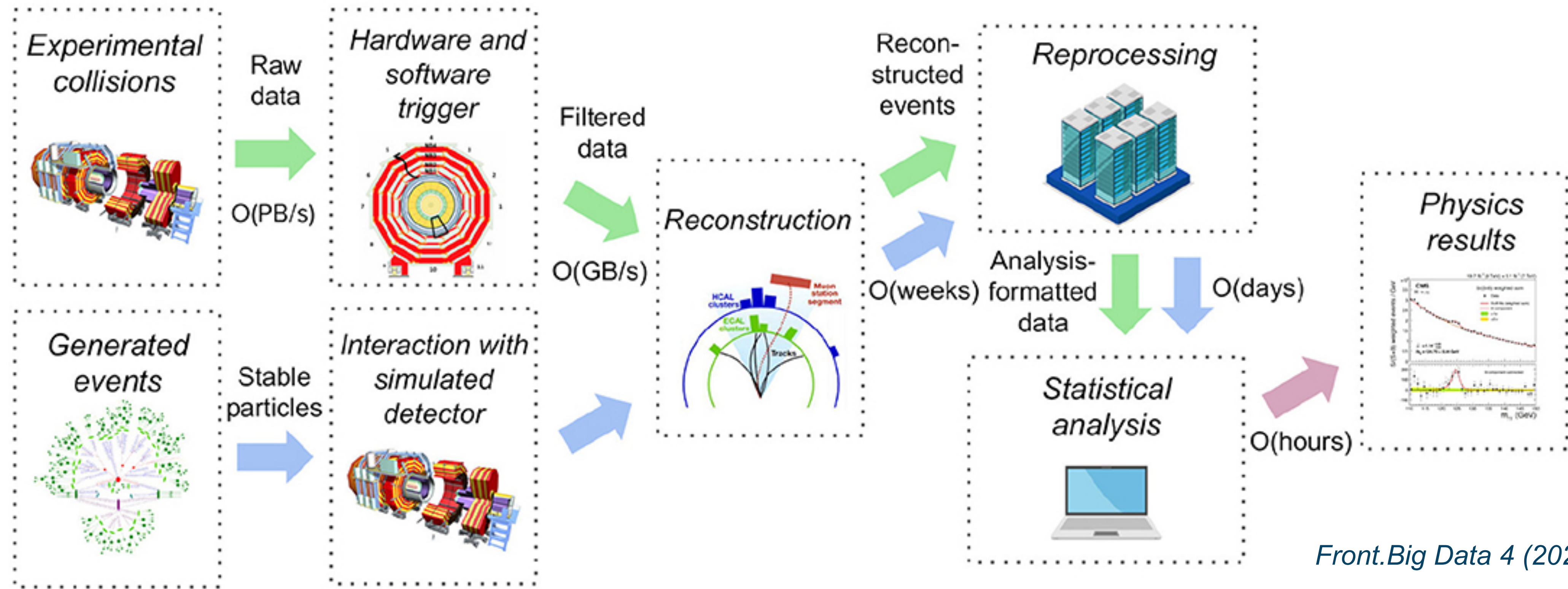
LHC, RHIC and EIC

- **Large Hadron Collider (LHC)**
 - ➡ proton+proton (p+p), lead+lead (Pb+Pb) collisions, etc
- **Relativistic Heavy Ion Collider (RHIC)**
 - ➡ proton+proton (p+p), gold+gold (Au+Au) collisions, etc
 - ➡ ended on Feb. 2026
- **Electron-Ion Collider (EIC)**
 - ➡ electron+proton (e+p), electron+gold (e+Au) collisions, etc
 - ➡ scheduled to begin data taking by ~2033



Nuclear and High-Energy Experiment Workflow

Detector design Experiment operation Data analysis Physics discovery

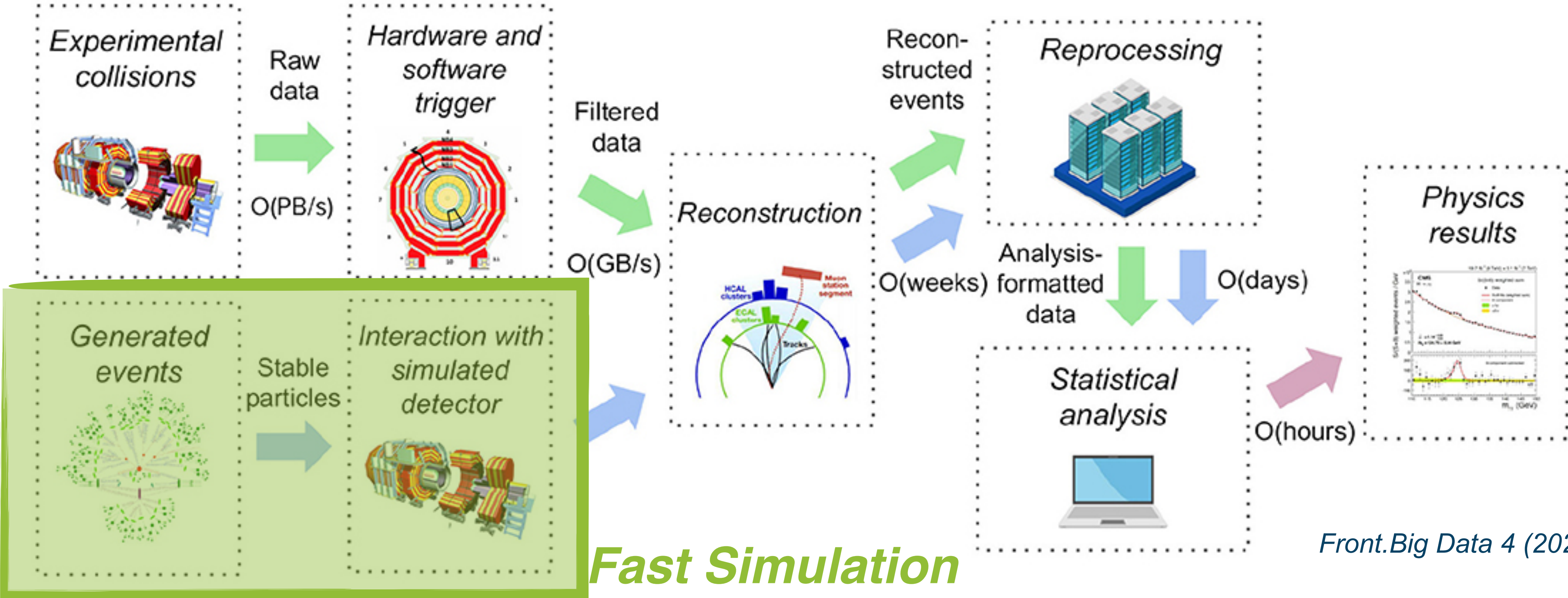


Front.Big Data 4 (2021) 661501

- Each stage of analysis requires substantial human effort, computing resources, and time
→ **Leverage AI/ML to accelerate, automate, and improve these processes**

Nuclear and High-Energy Experiment Workflow

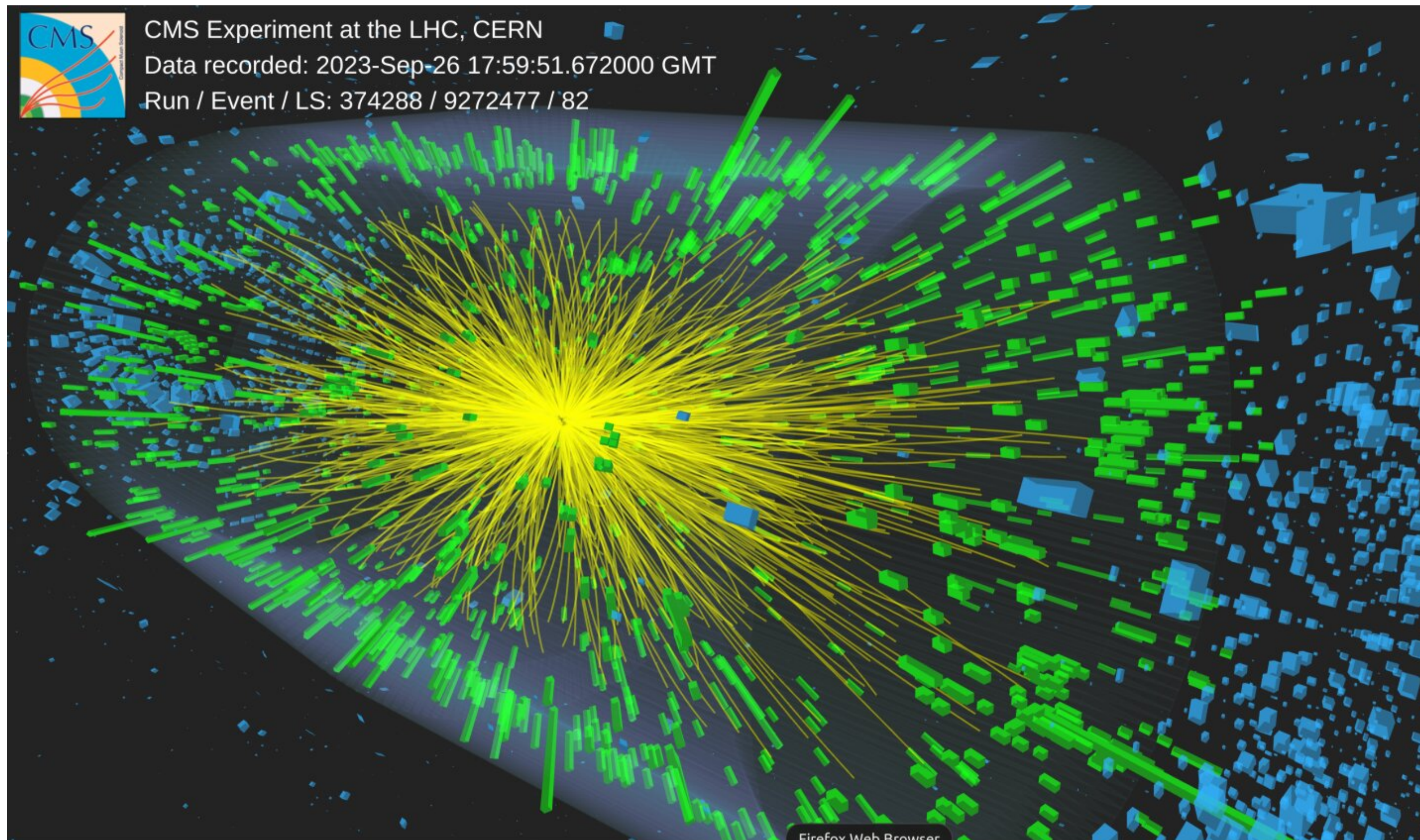
Detector design Experiment operation Data analysis Physics discovery



Front.Big Data 4 (2021) 661501

Simulations of Relativistic Heavy Ion Collisions

- **$O(1000)$ particles** in single nuclear collision event + **thousands shower steps** per particle
➔ Simulation of the interaction of particles with detectors is **high complexity and computationally intensive work**



CMS Pb+Pb collision
experimental data
event display

The sPHENIX Detector

1.4T Superconducting Magnet

Outer Hadronic Calorimeter

Inner Hadronic Calorimeter

Electromagnetic Calorimeter

MVTX

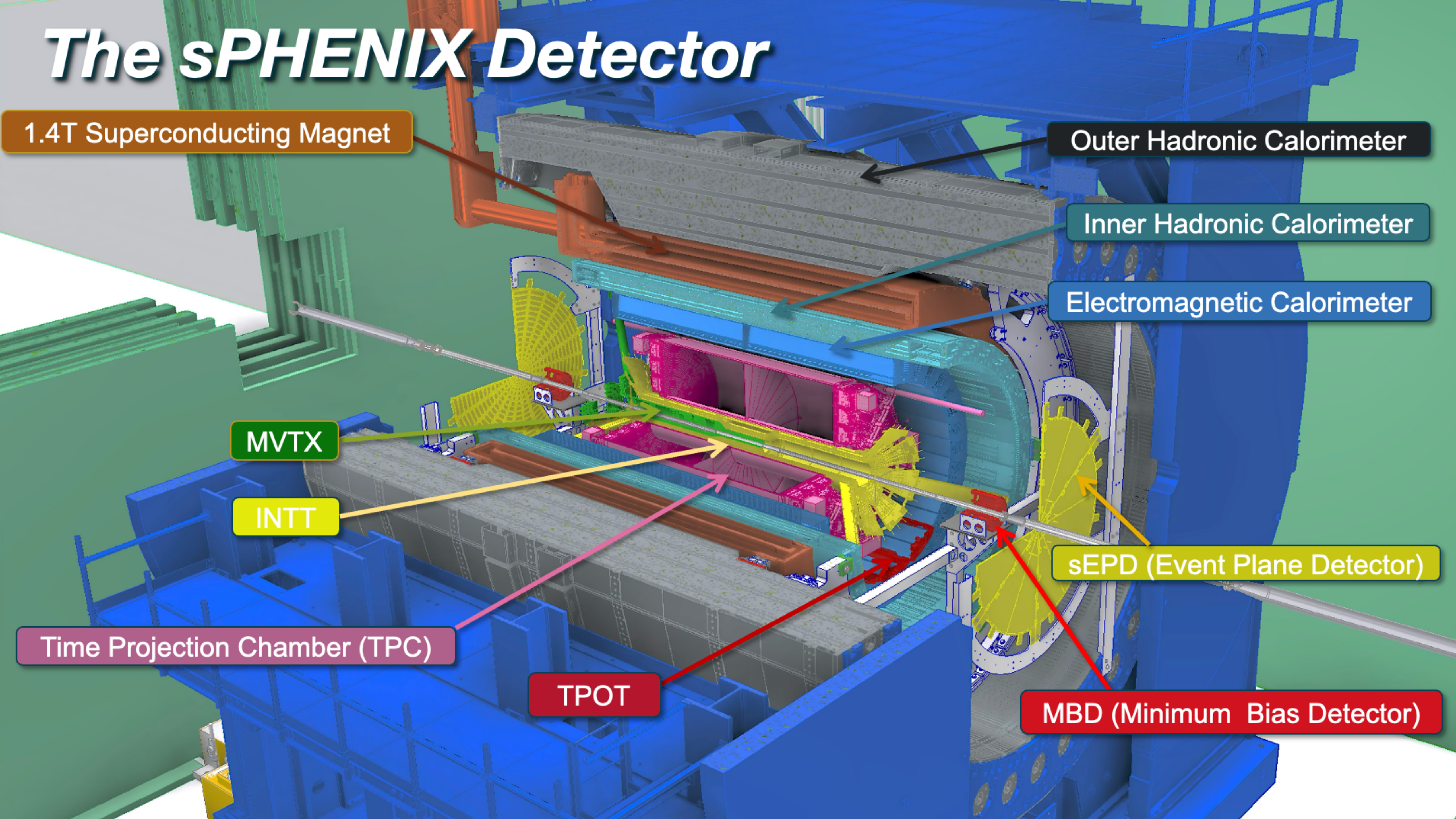
INTT

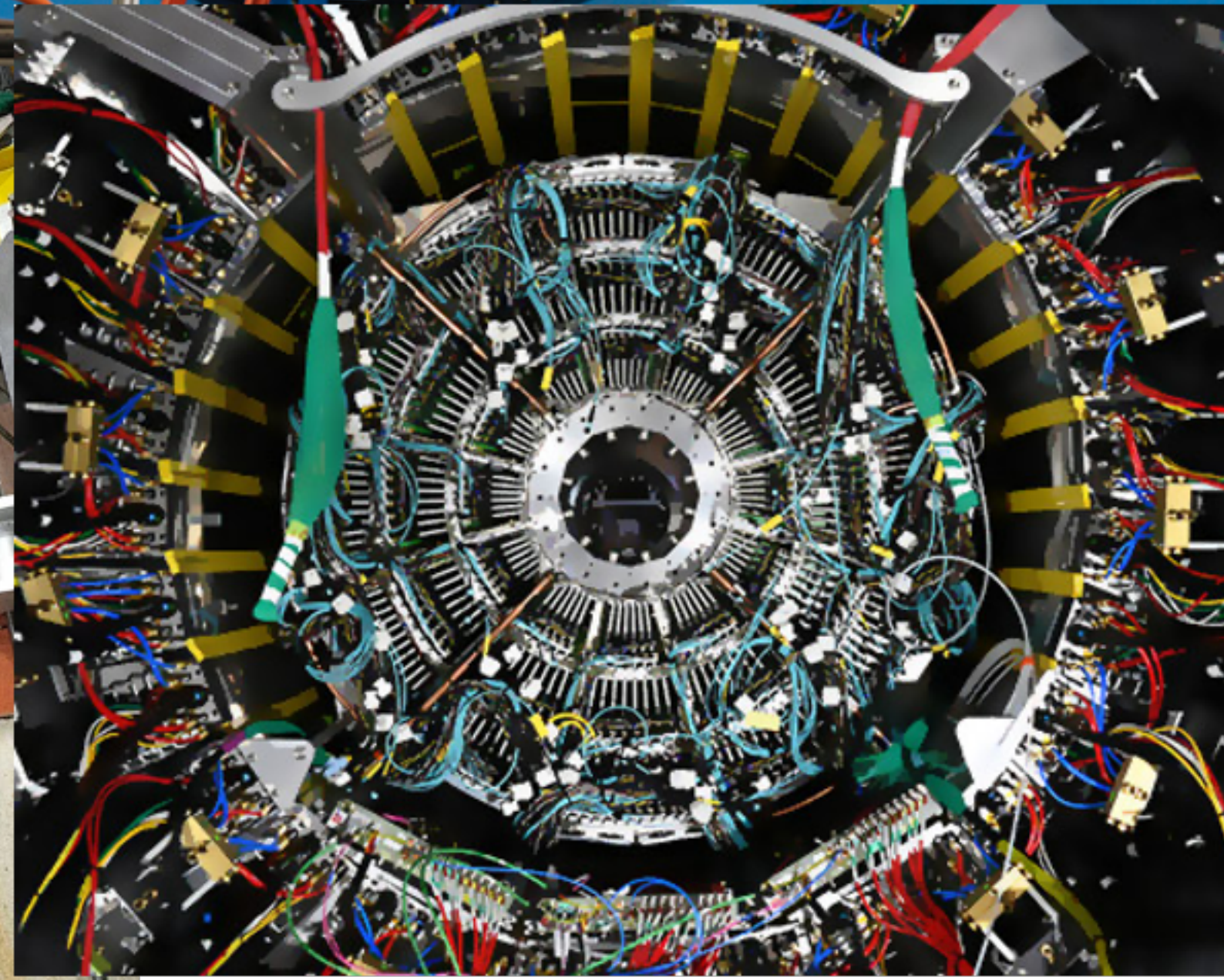
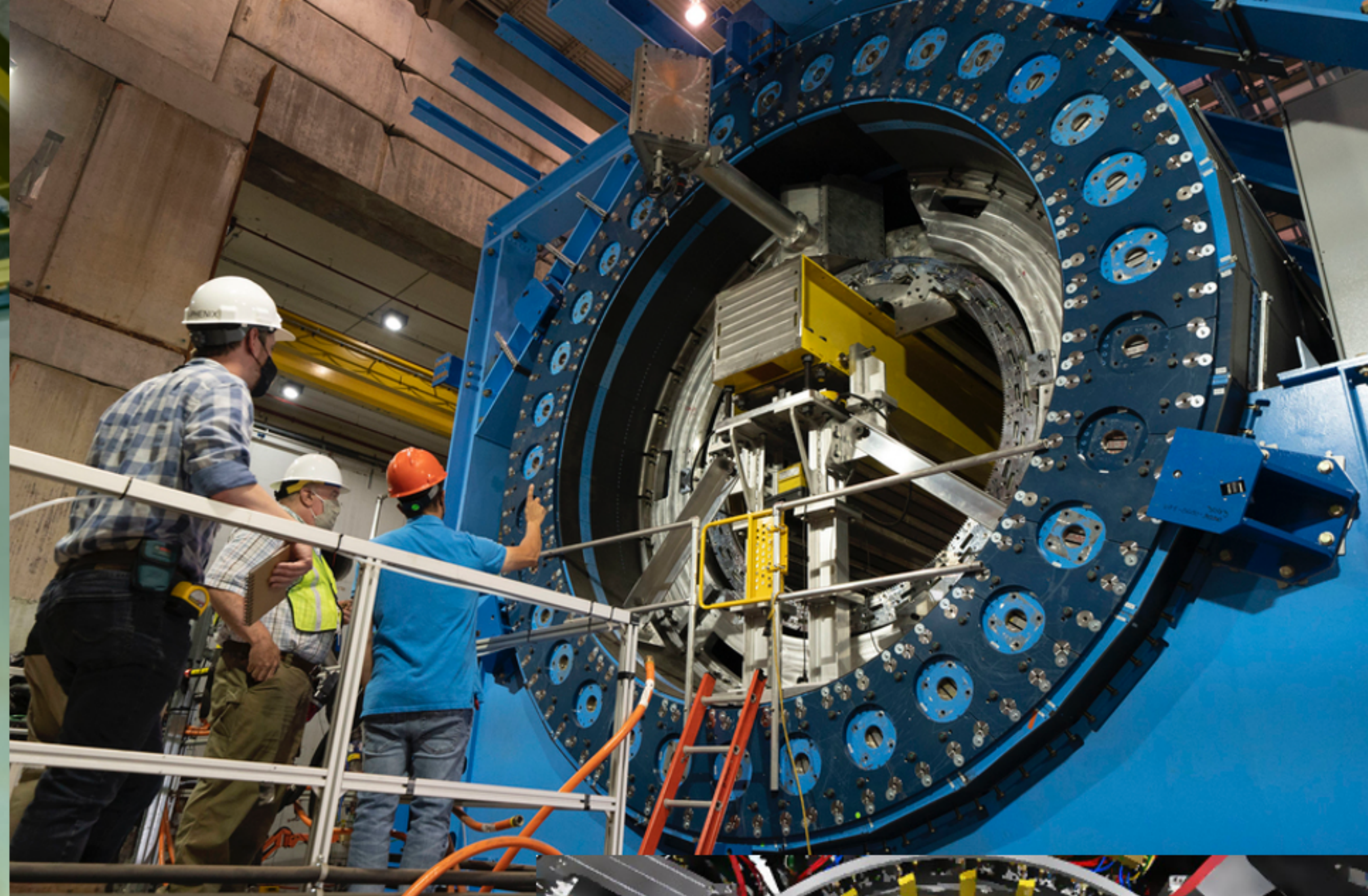
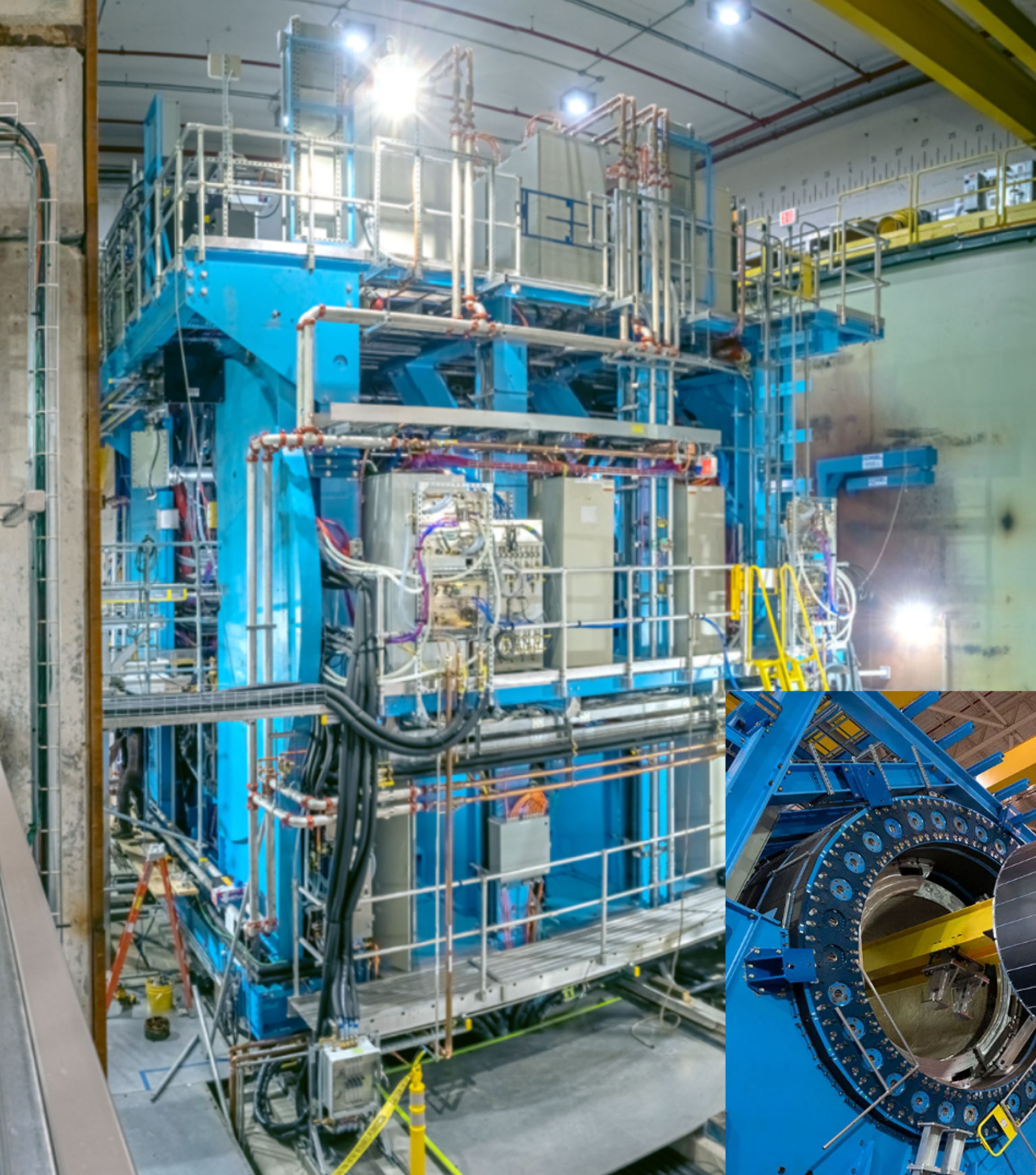
Time Projection Chamber (TPC)

TPOT

sEPD (Event Plane Detector)

MBD (Minimum Bias Detector)





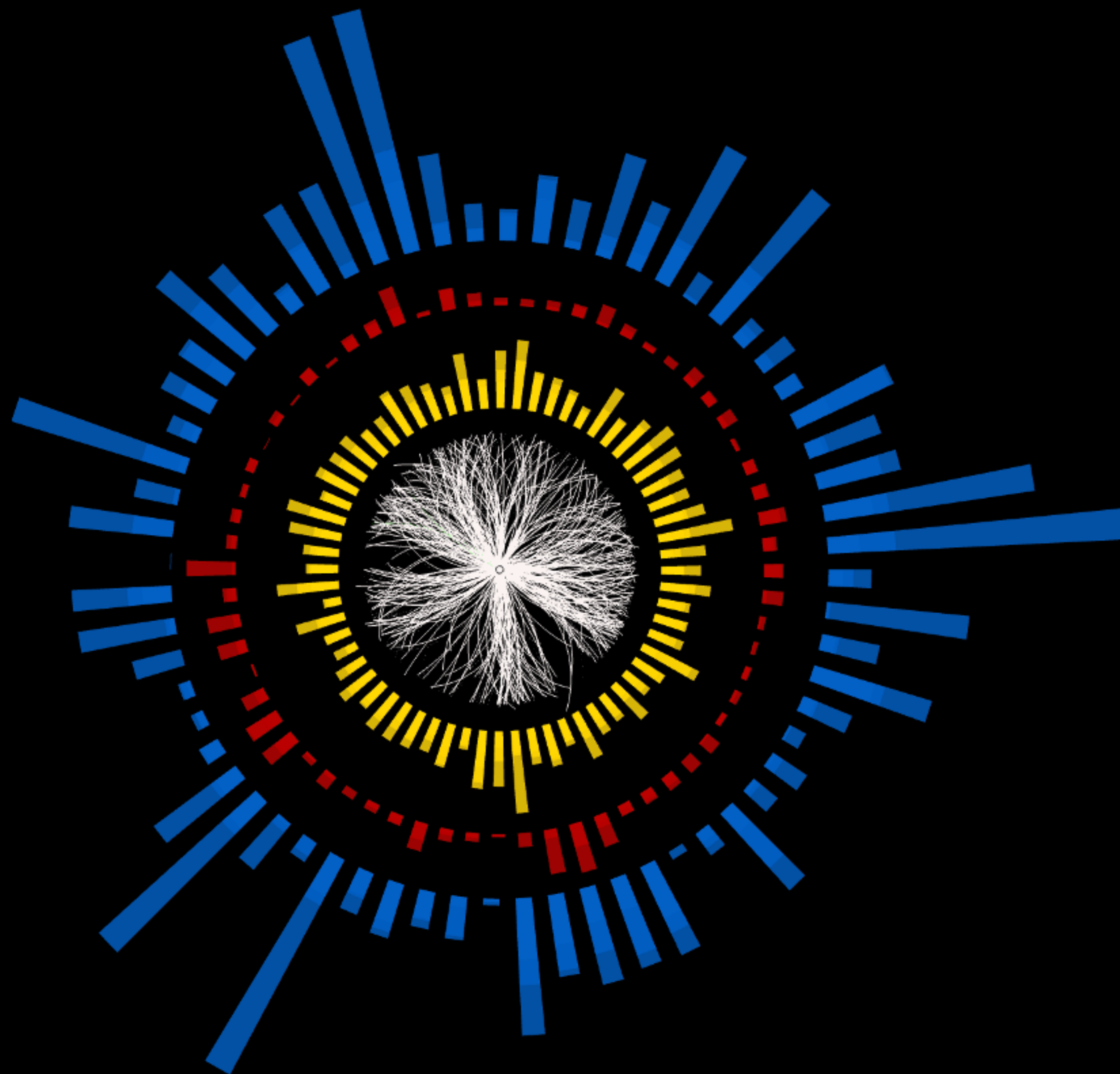
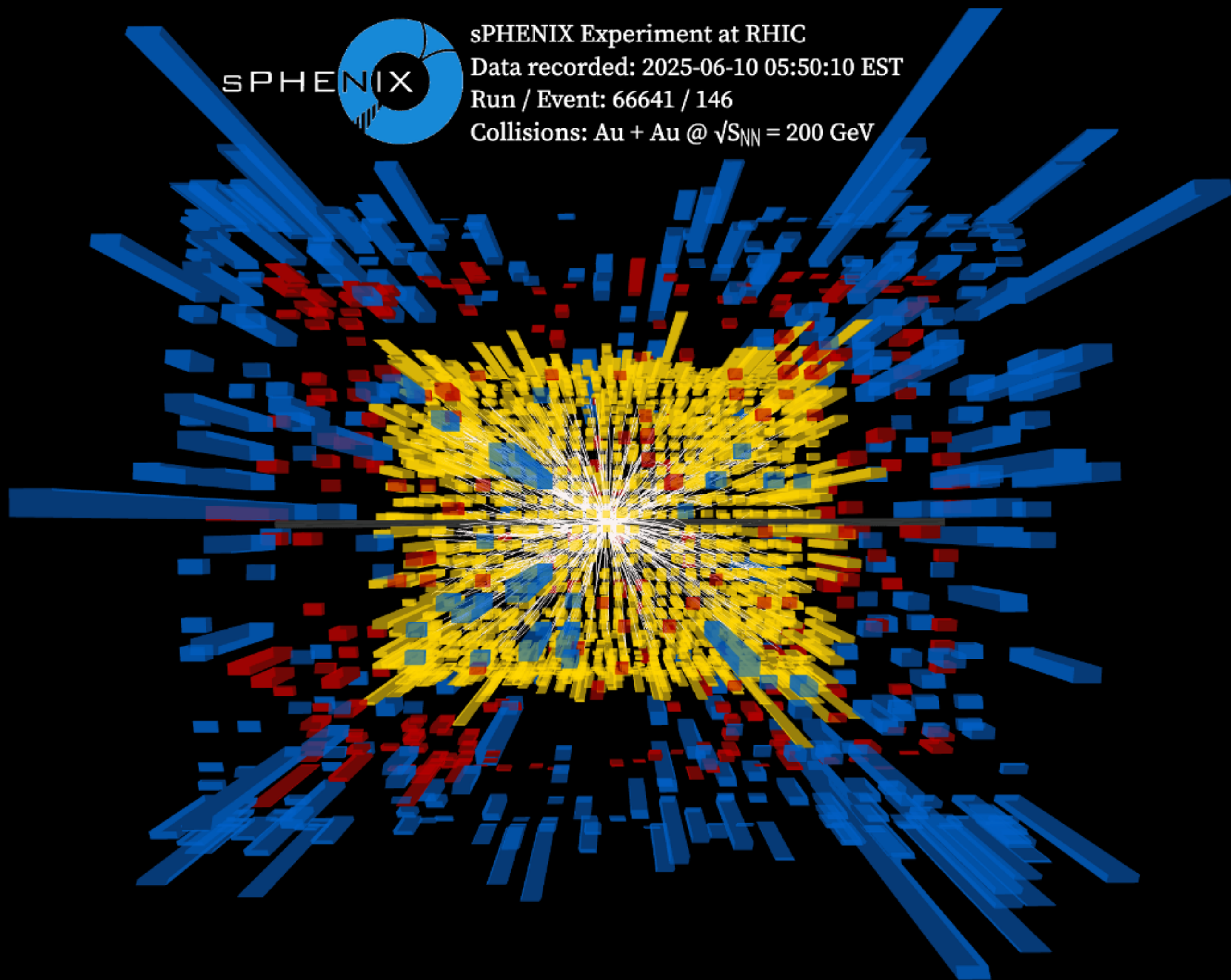


sPHENIX Experiment at RHIC

Data recorded: 2025-06-10 05:50:10 EST

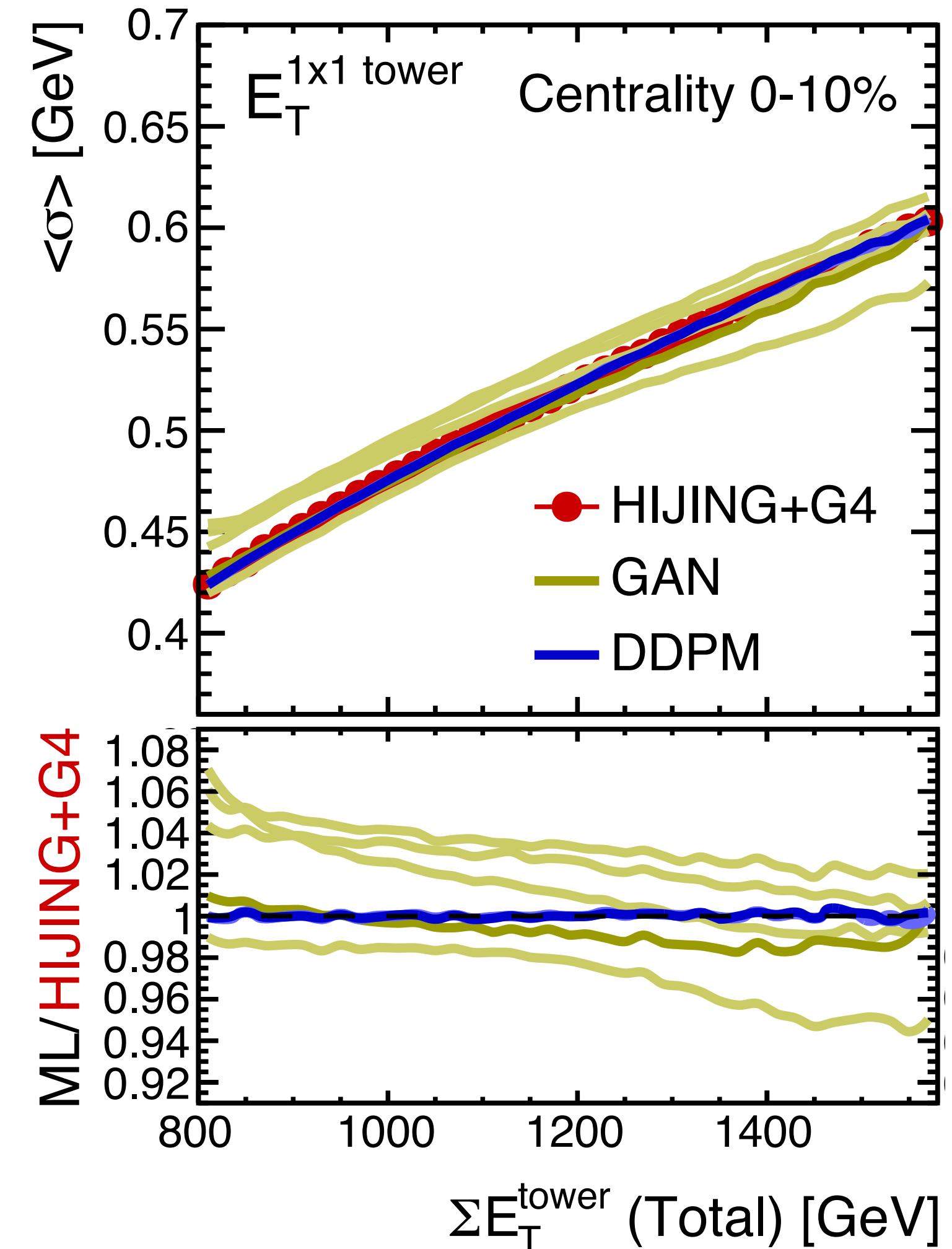
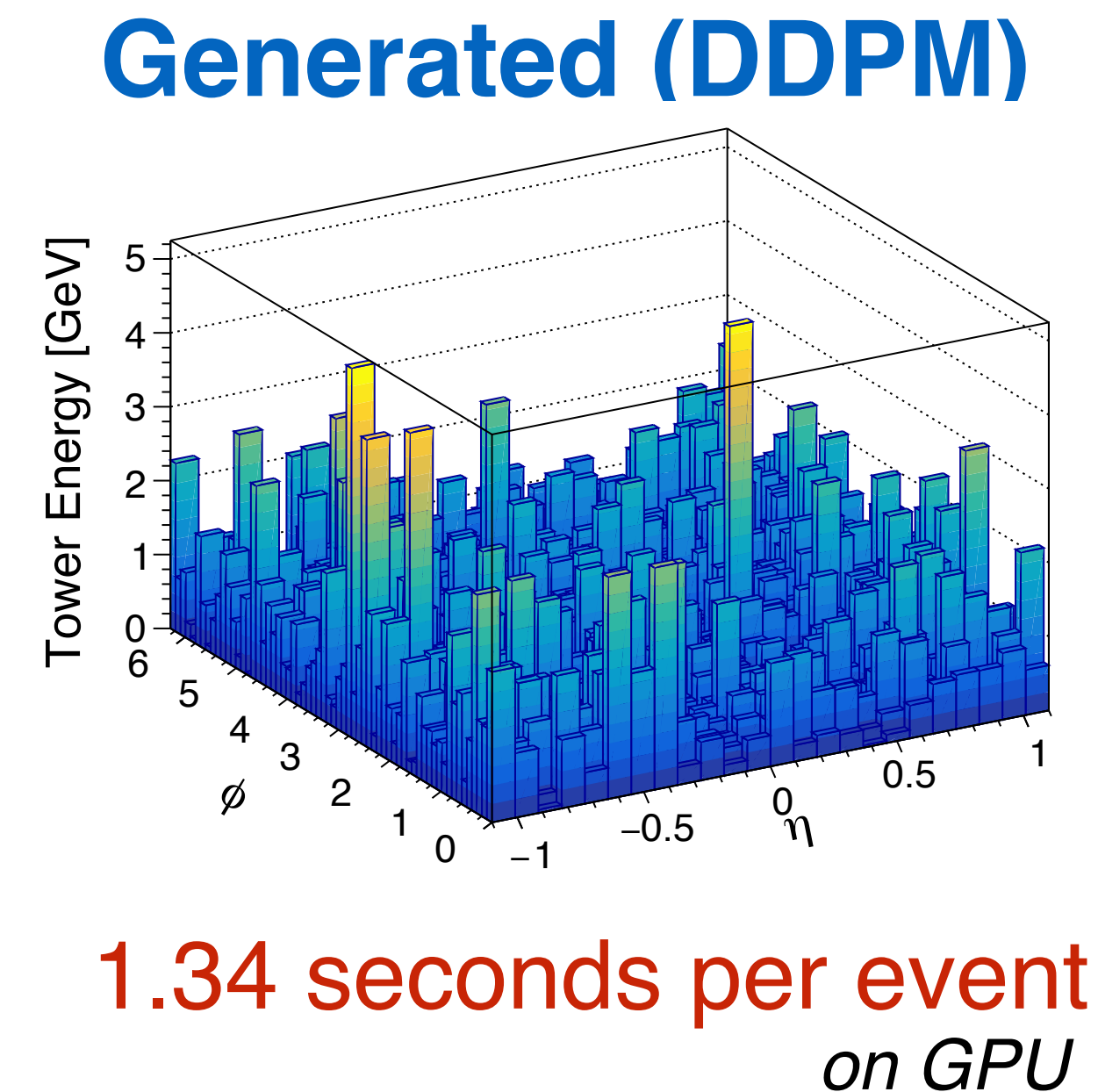
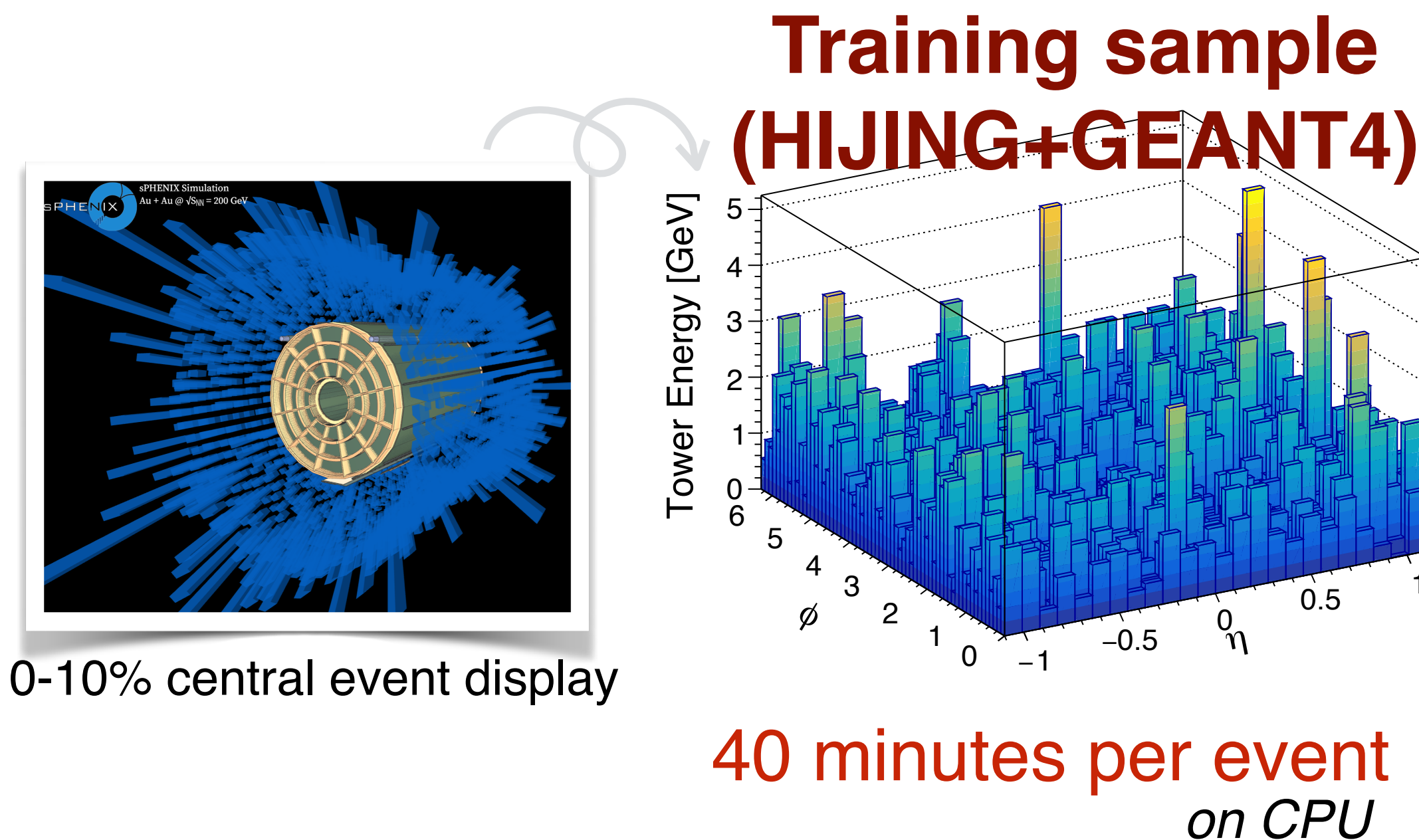
Run / Event: 66641 / 146

Collisions: Au + Au @ $\sqrt{s_{NN}} = 200$ GeV



Whole-Event and Full-Calorimeter Simulation

- Generate high fidelity “particle production + full detector simulation” data in *central heavy ion collisions* using Denosing Diffusion Probabilistic Model (**DDPM**) [PRC 110 \(2024\) 034912](#)

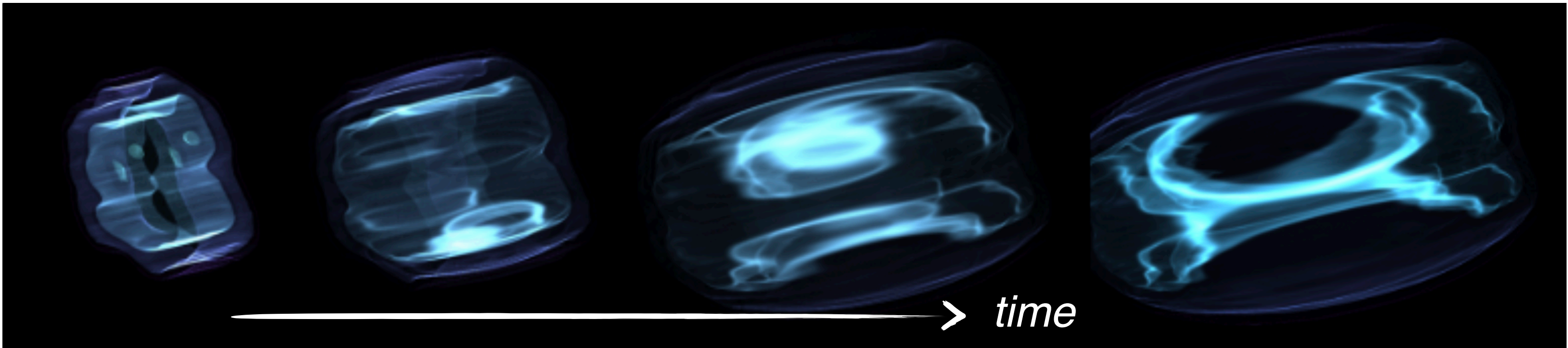


- DDPM** provide a **speedup of 1800 times!** (on GPU)
➔ at least 100 times faster, considering a 32-core CPU equivalent to a GPU
- Great stability and accuracy of **DDPM** within sub-percent level

Hydrodynamic Evolution Simulation

- Full *spatiotemporal hydrodynamic medium (Quark Gluon Plasma) evolution simulations* are computationally very demanding

QGP 3+1D Energy Density Evolution

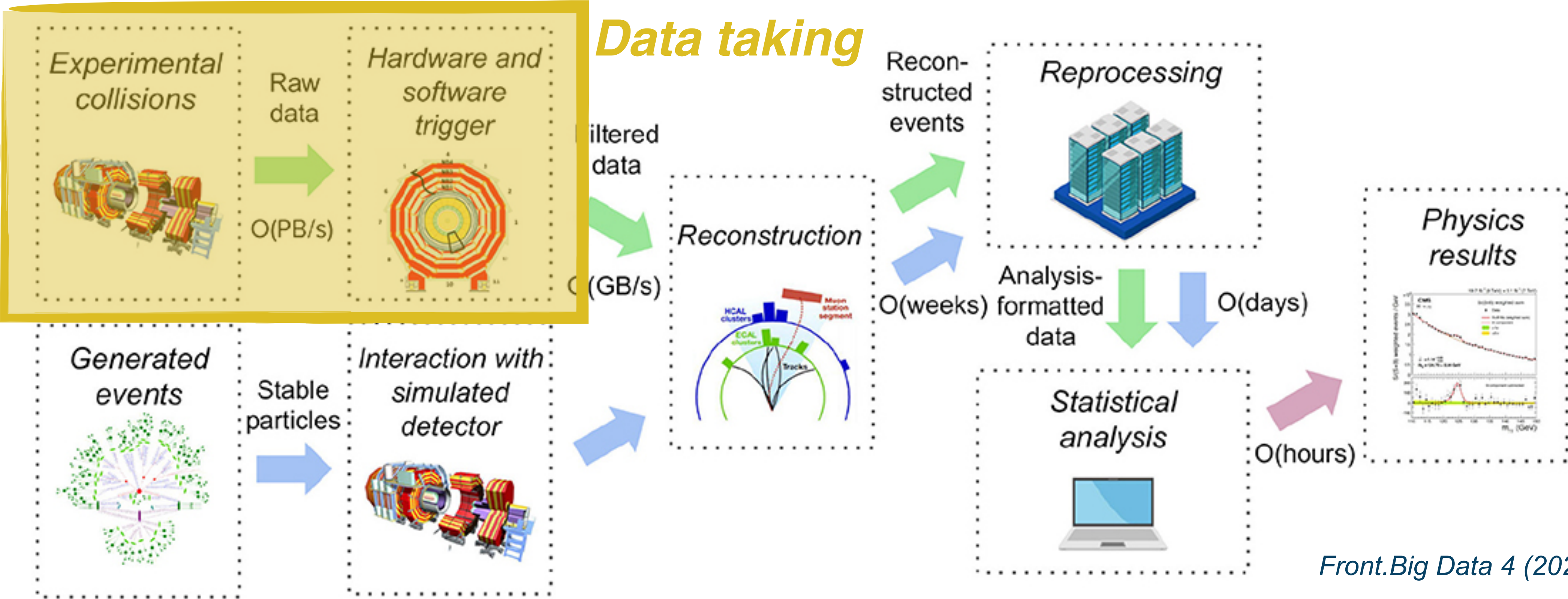


- **Fourier Neural Operator (FNO)**-based model developed: **TC-QGP** to predict the full time evolution of 3D QGP fields from the initial time 3D fields.
 - **$O(100,000)$ speedup with a few percent accuracy**

	2+1D	3+1D
Simulation time per event with 1 CPU	~10 minutes	~20 hours
TC-QGP	0.03 seconds	0.2 seconds

Nuclear and High-Energy Experiment Workflow

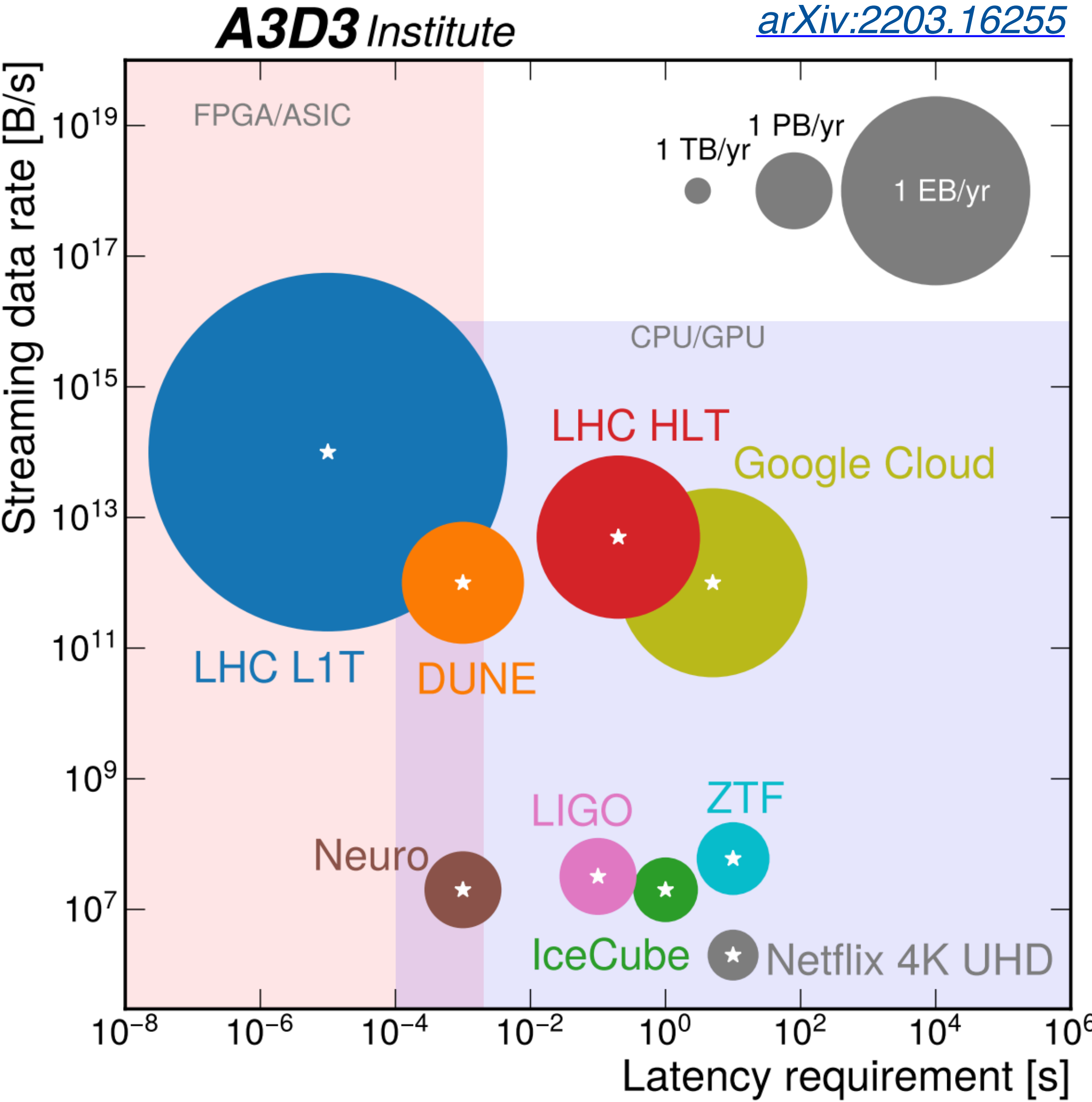
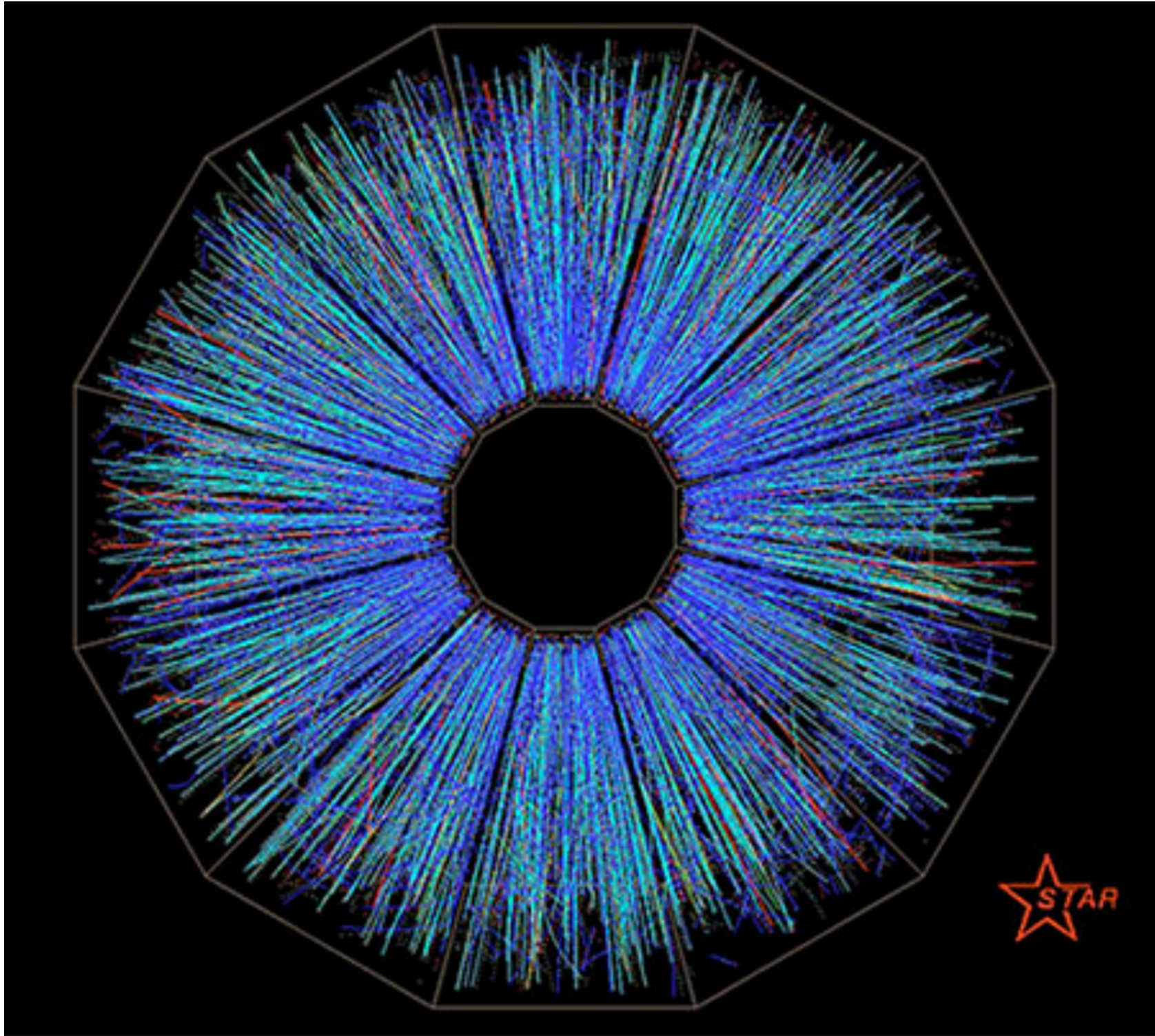
Detector design Experiment operation Data analysis Physics discovery



Front.Big Data 4 (2021) 661501

High Energy Collision Data Taking

Collision rates 50kHz - 3MHz
Raw data: O(TB/s) (future detectors O(PB/s))

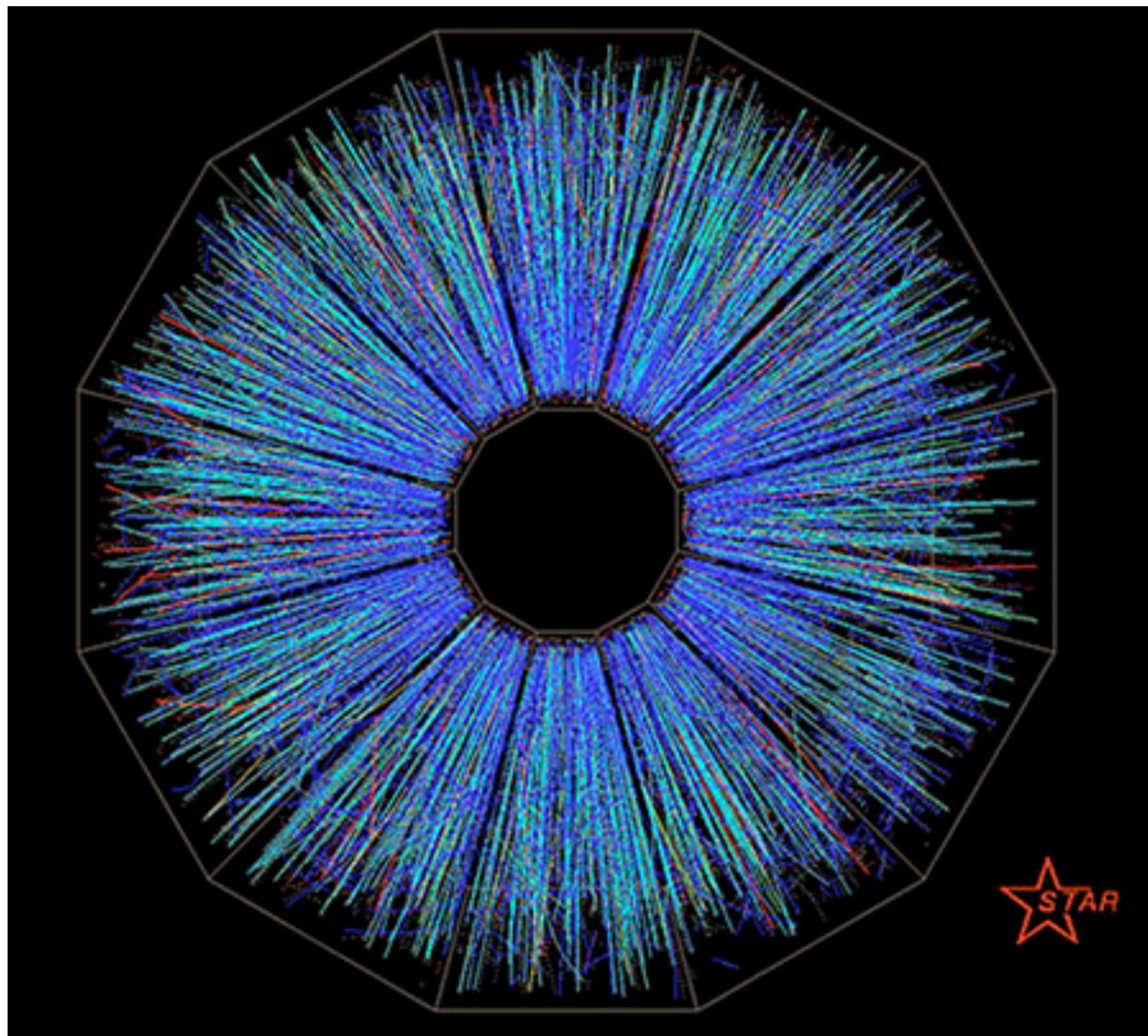


High Energy Collision Data Taking

Collision rates 50kHz - 3MHz

Raw data: O(TB/s) (future detectors O(PB/s))

MLP or GNN tracking on FPGA
JINST 19 (2024) C02066,
Front.Big Data 5 (2022) 828666
arXiv:2112.02048



Triggering

Is it interesting event? → if yes, save!
(Real-time filtering)

Streaming readout ← next generation
*Let's save everything by **compressing** data!*

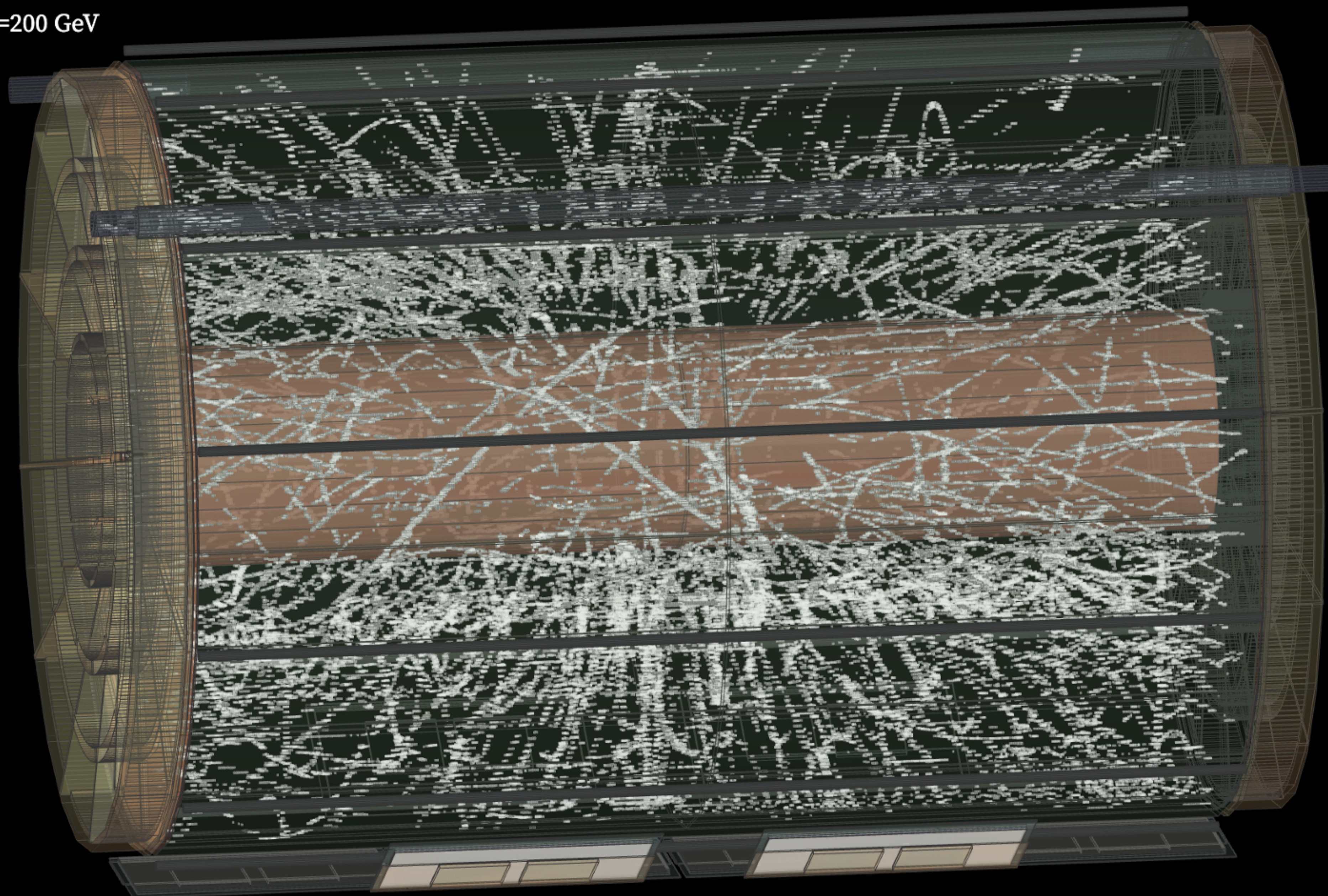


sPHENIX Time Projection Chamber

100 Hz ZDC, MBD Prescale: 2, HV: 4.45 kV GEM, 45 kV CM, X-ing Angle: 2 mrad

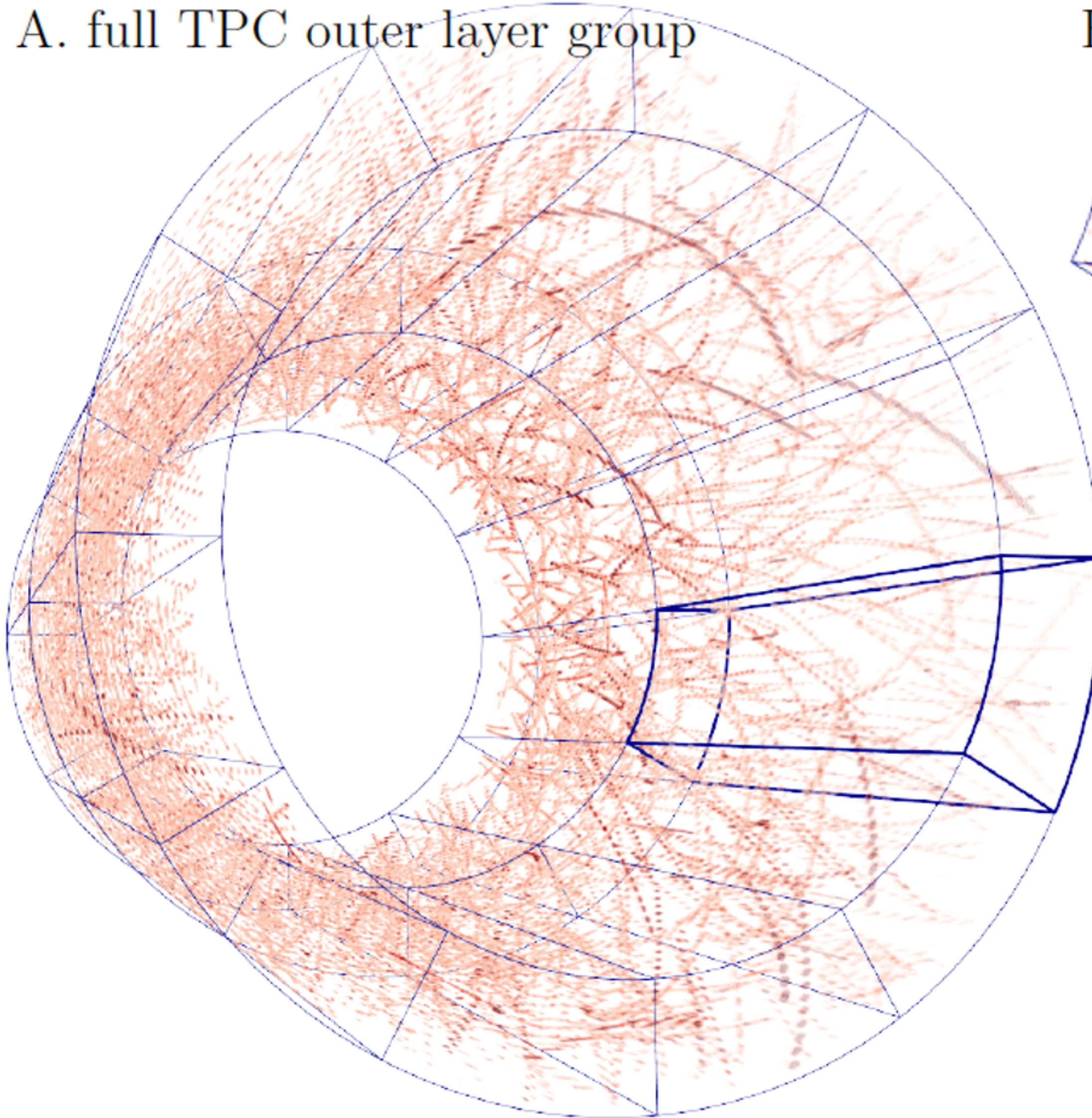
2023-06-23, Run 10931 - EBDC03 reference frame 89

Au+Au $\sqrt{s_{NN}}=200$ GeV

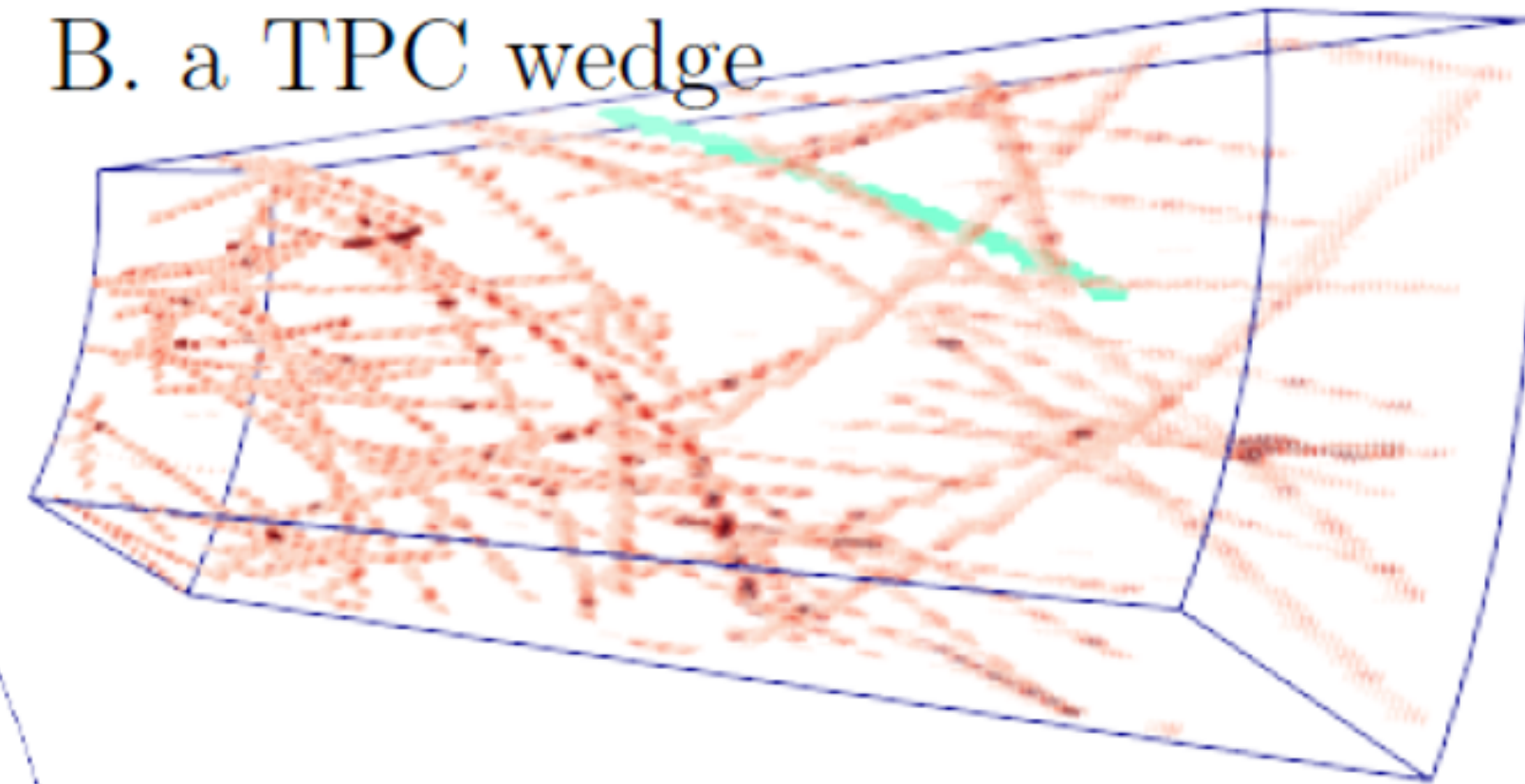


sPHENIX TPC Event Display for Heavy Ion Collisions

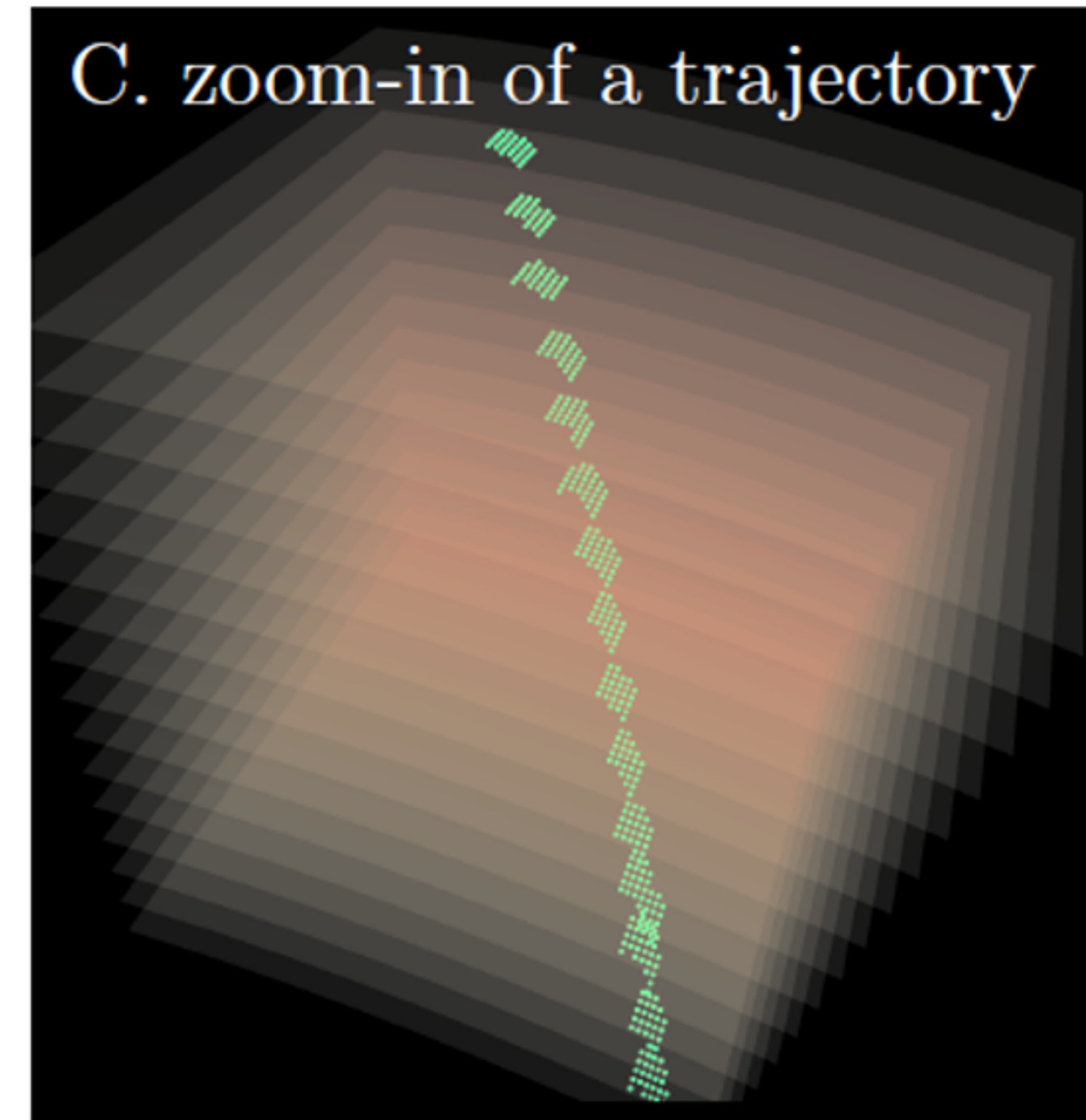
A. full TPC outer layer group



B. a TPC wedge



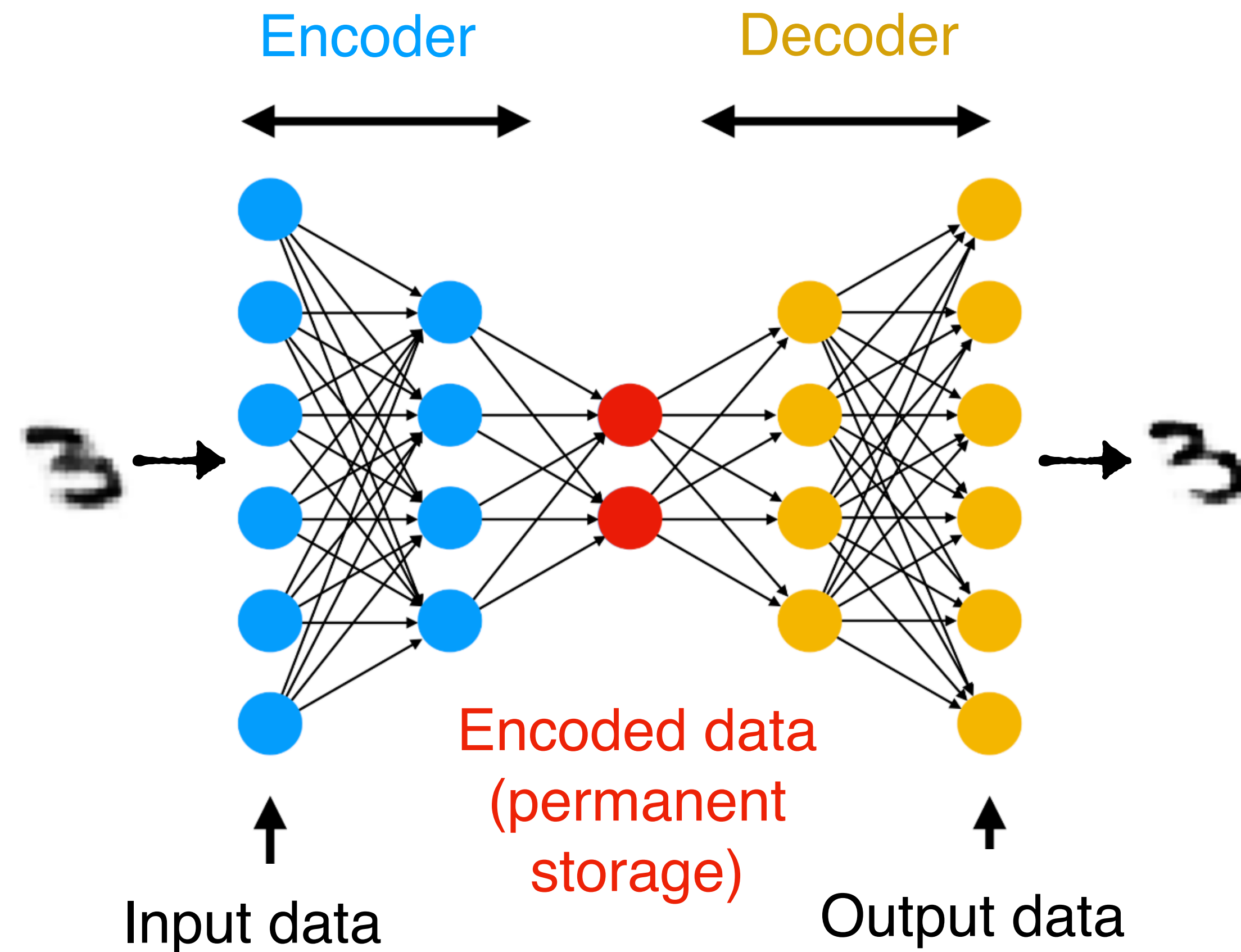
C. zoom-in of a trajectory



Data Compression using AutoEncoder

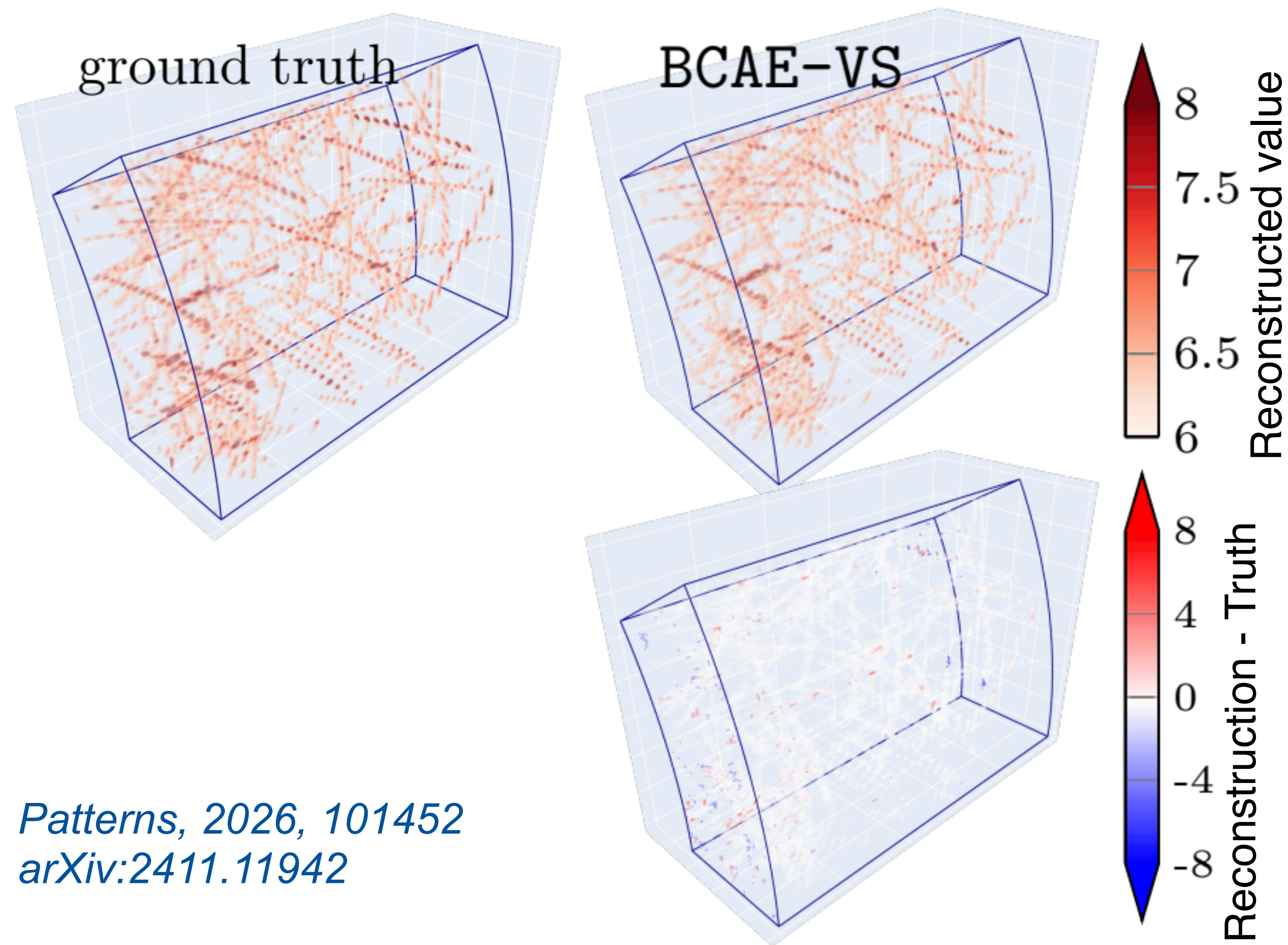
- Data Compression

➔ Record compact **encoded** data during data taking, **store them permanently**, and **decode** them offline for reconstruction

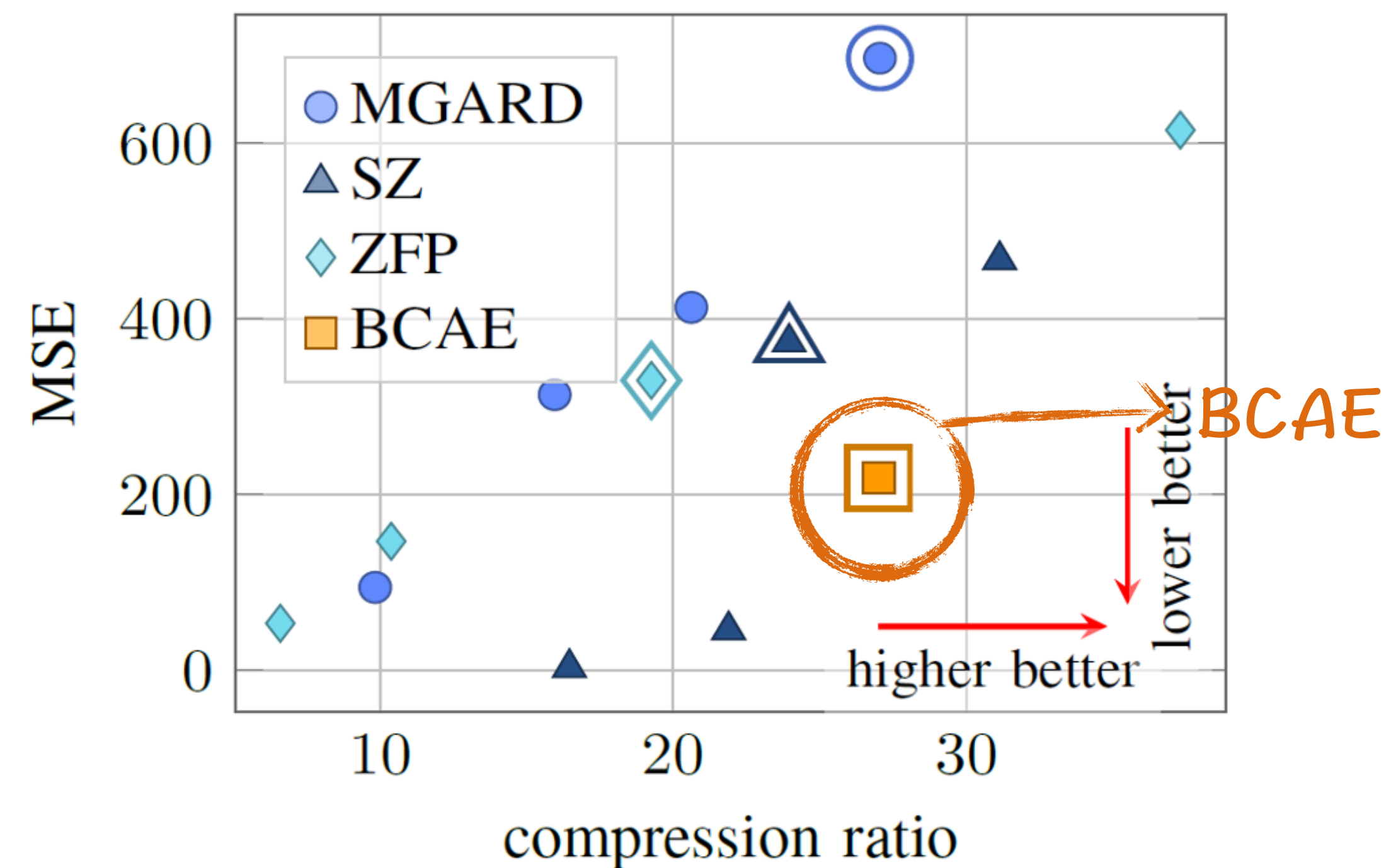


Data Compression using AutoEncoder

- *BCAE (Bicephalous Convolutional Autoencoder) model*



Patterns, 2026, 101452
arXiv:2411.11942

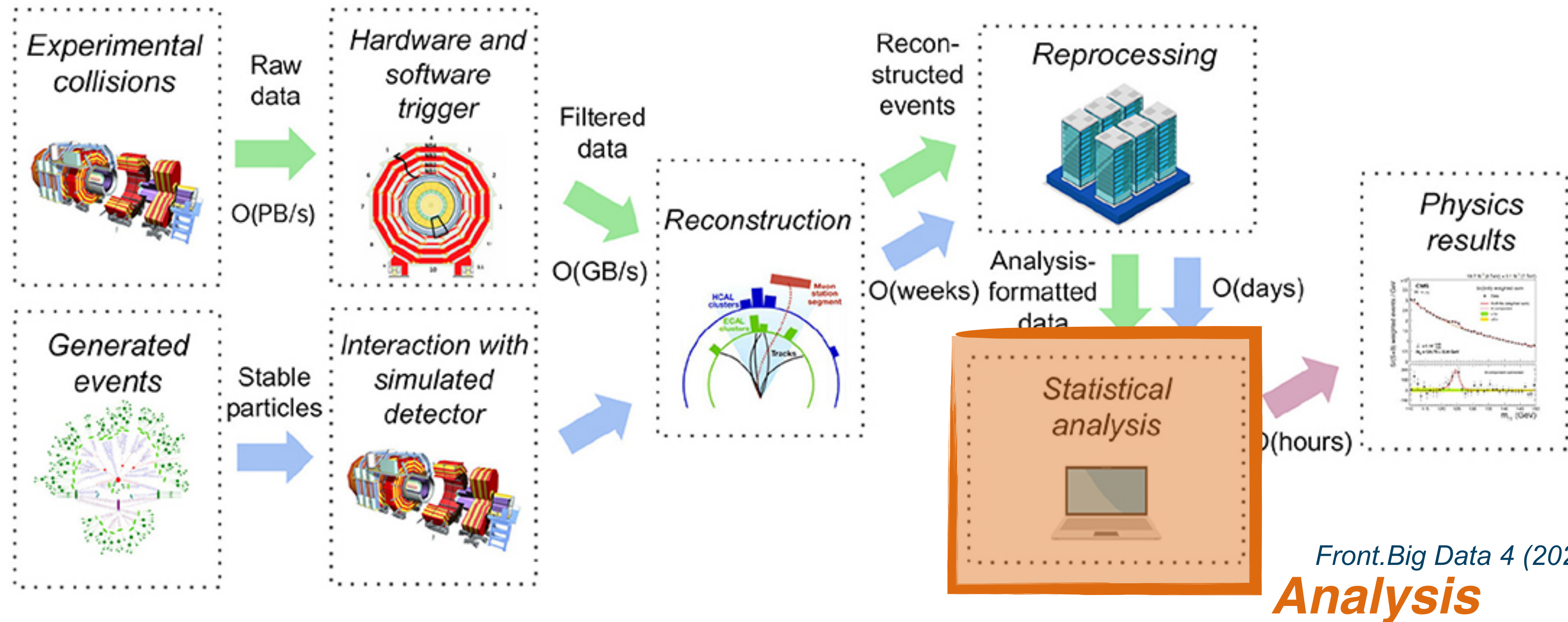


BCAE outperforms other off-the-shelf lossy compression algorithms

BCAE-VS: Great performance for sparse data (great prospects towards EIC!)

Nuclear and High-Energy Experiment Workflow

Detector design Experiment operation Data analysis Physics discovery

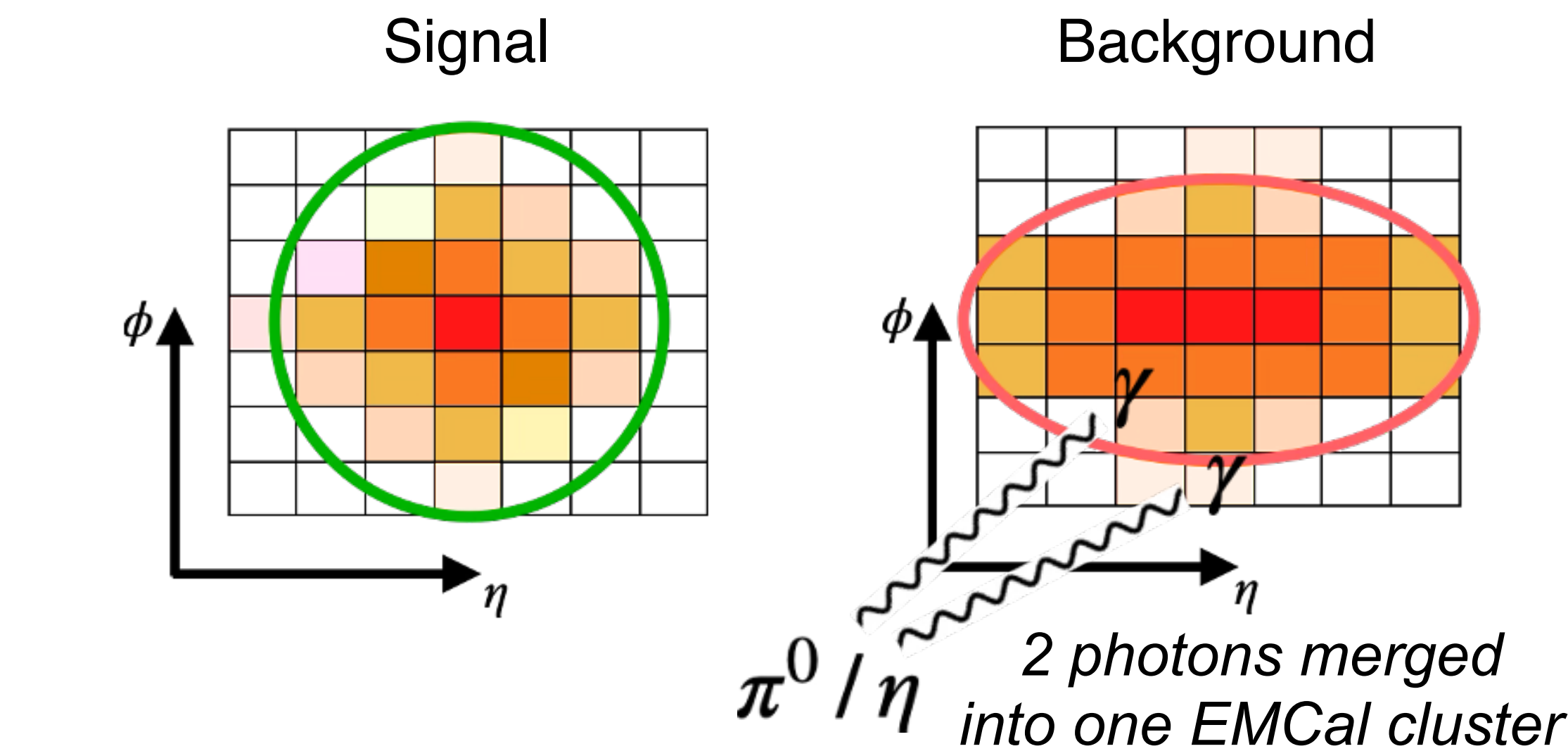


- Each stage of analysis requires substantial human effort, computing resources, and time
→ *Leverage AI/ML to accelerate, automate, and improve these processes*

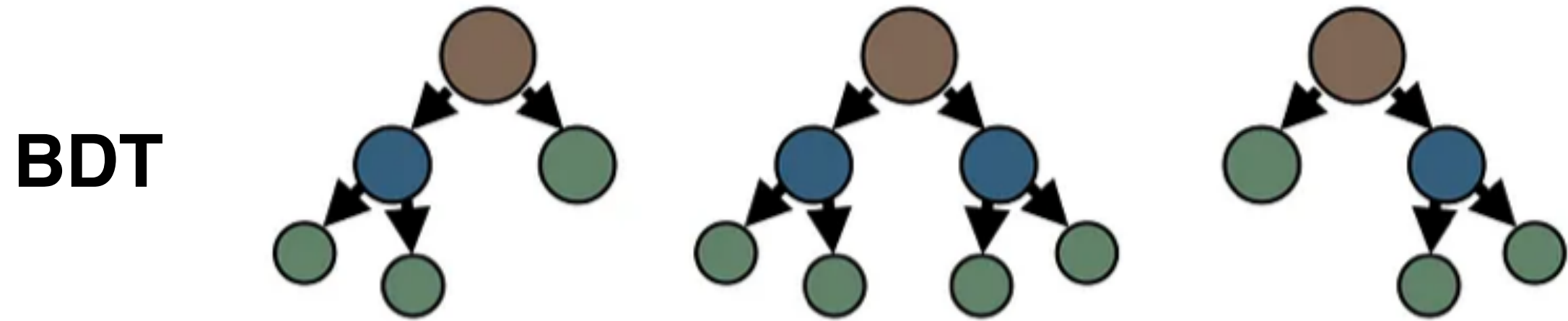
Classification: Signal vs. Background

[arXiv:2607.03892](https://arxiv.org/abs/2607.03892)

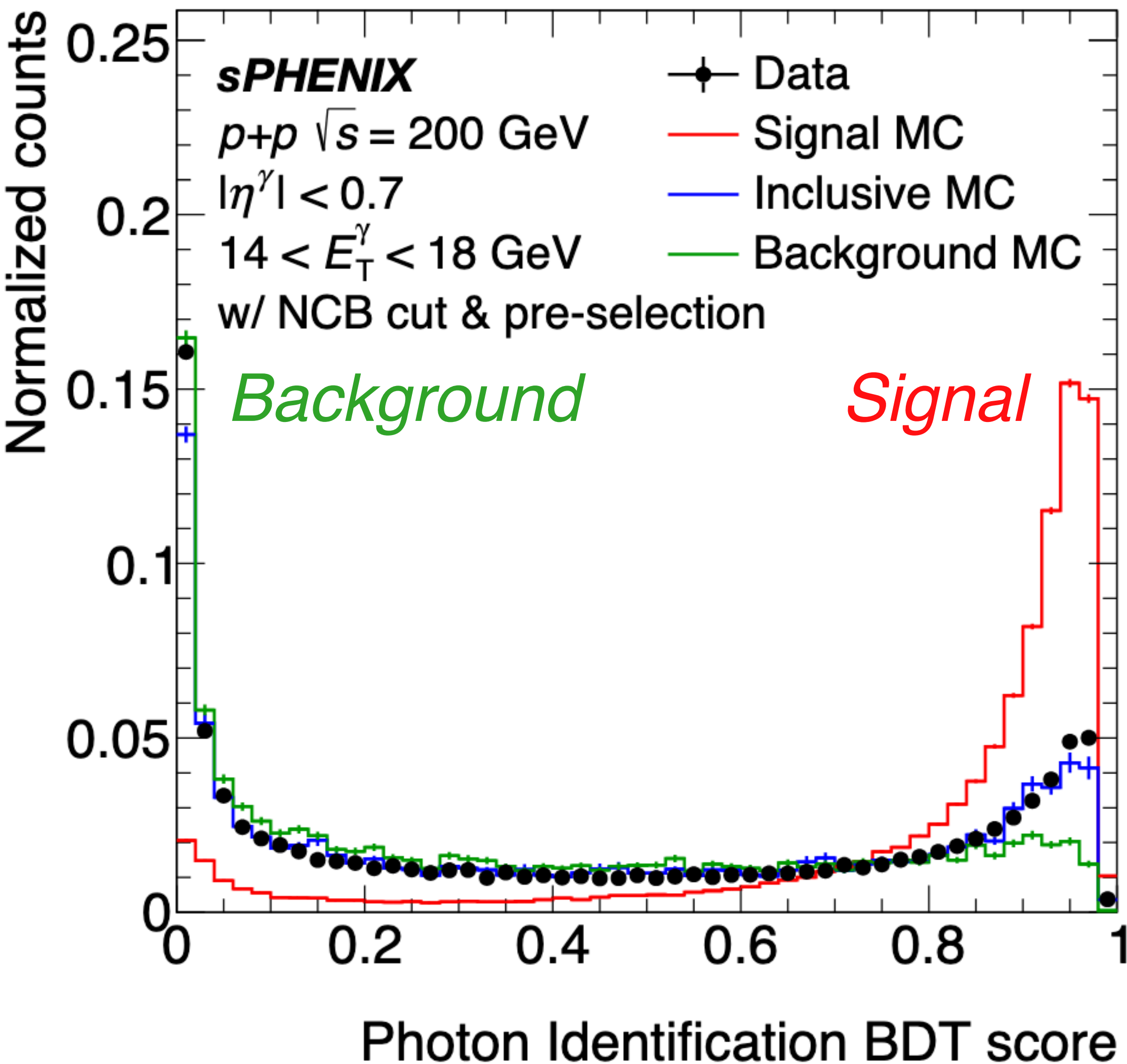
- **Boosted Decision Trees (BDT):** Multivariate method for signal-background classification



Input: Shower Shape Variables



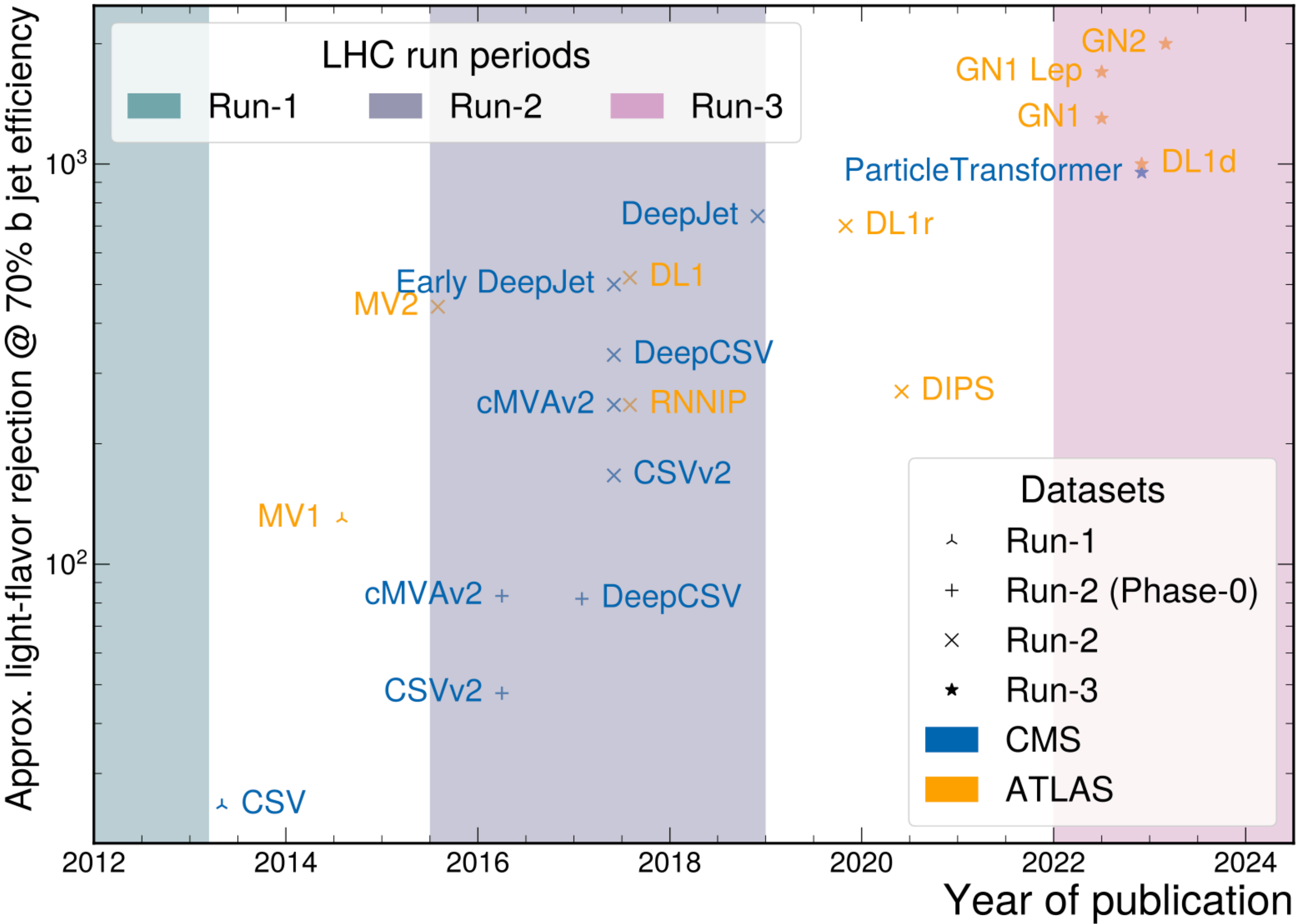
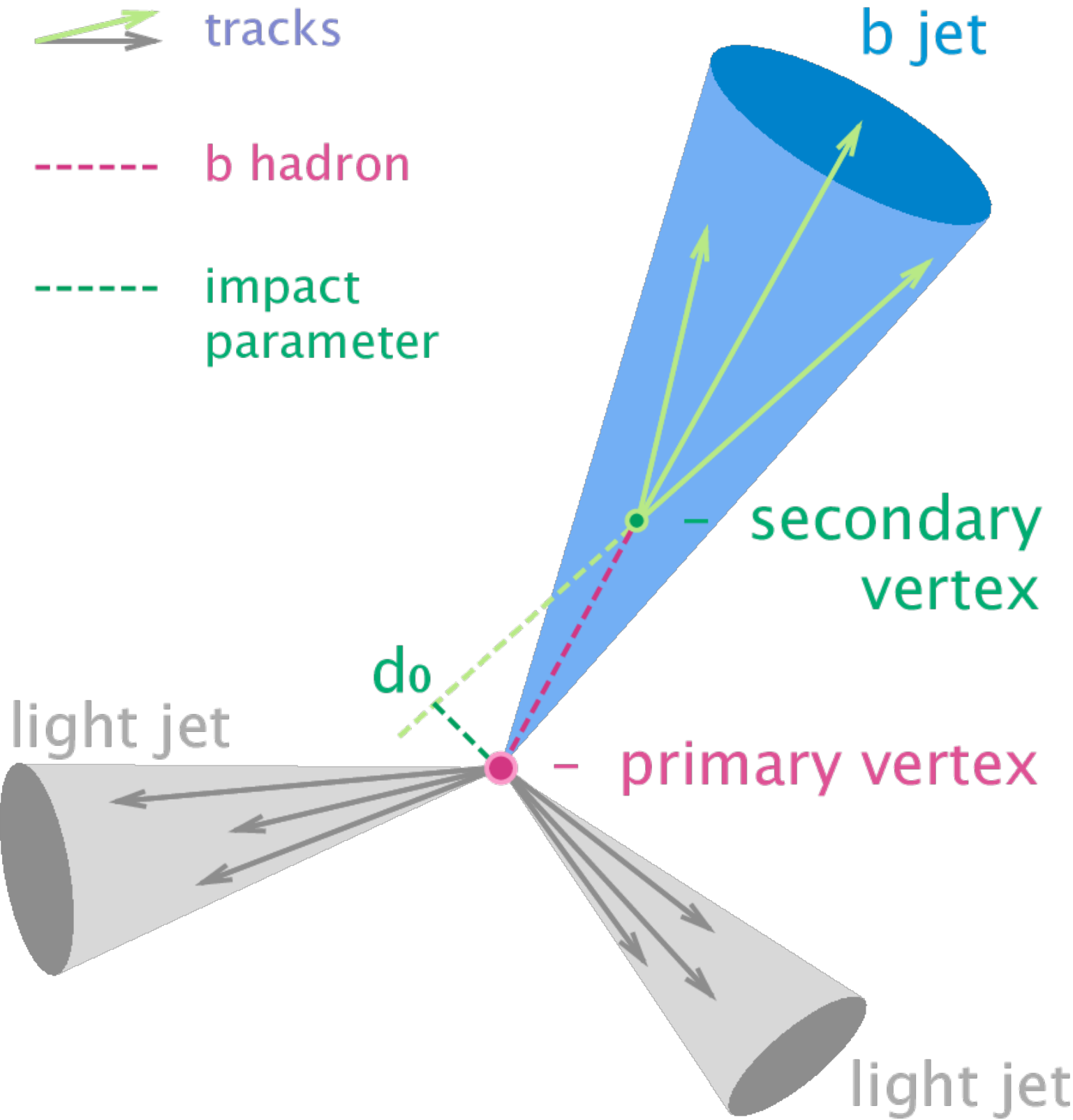
Output: Signal or Background?



Classification: Heavy Flavor Jet Tagging

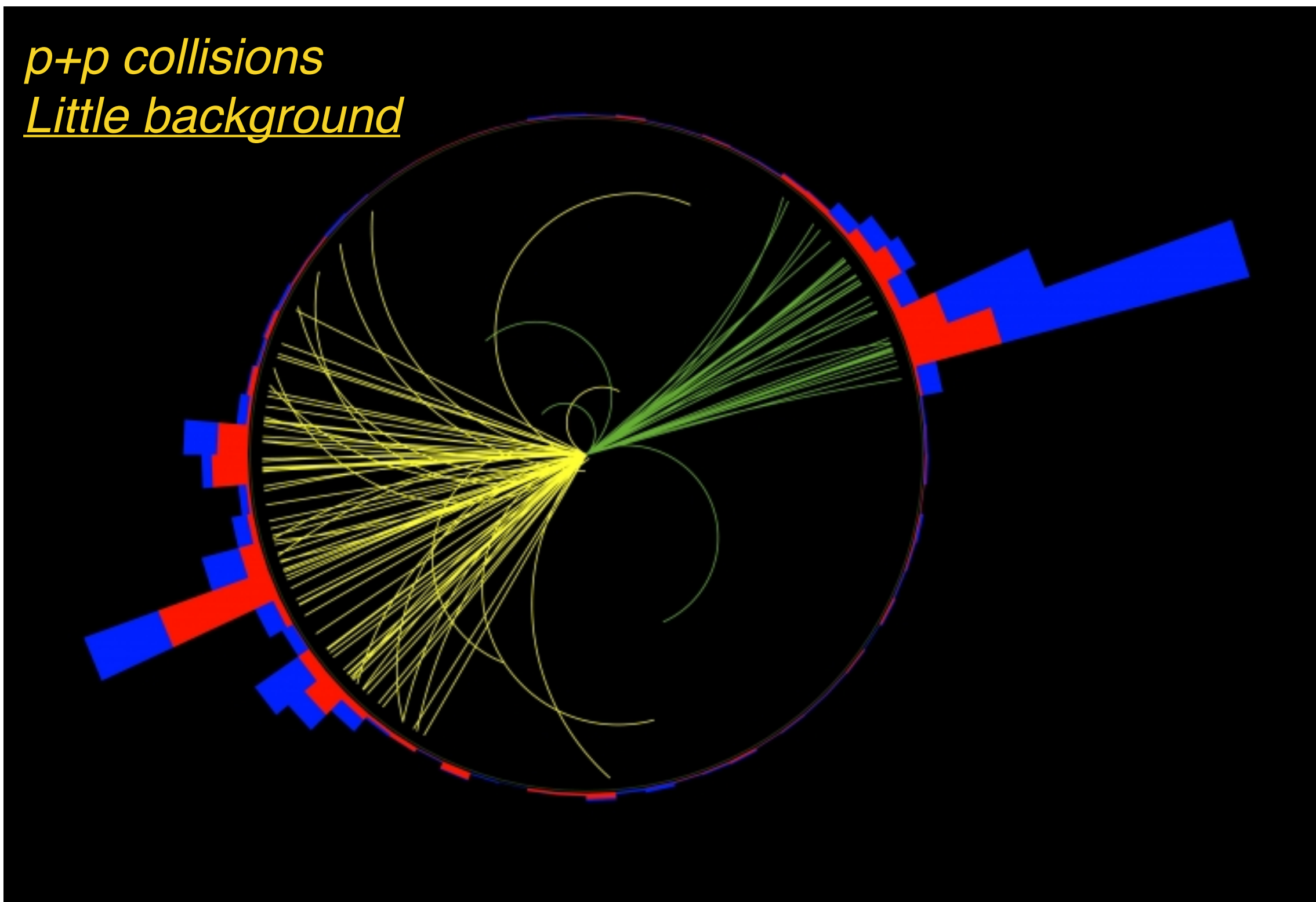
- Significant evolution on **Jet Flavor tagging** using various ML models

[Eur.Phys.J.ST 233 \(2024\) 15-16, 2657-2686](#)



Background in Jet Events in Heavy Ion Collisions

- **Signals (e.g. jet) lie on the significant underlying events (background) in heavy ion collisions**



sPHENIX

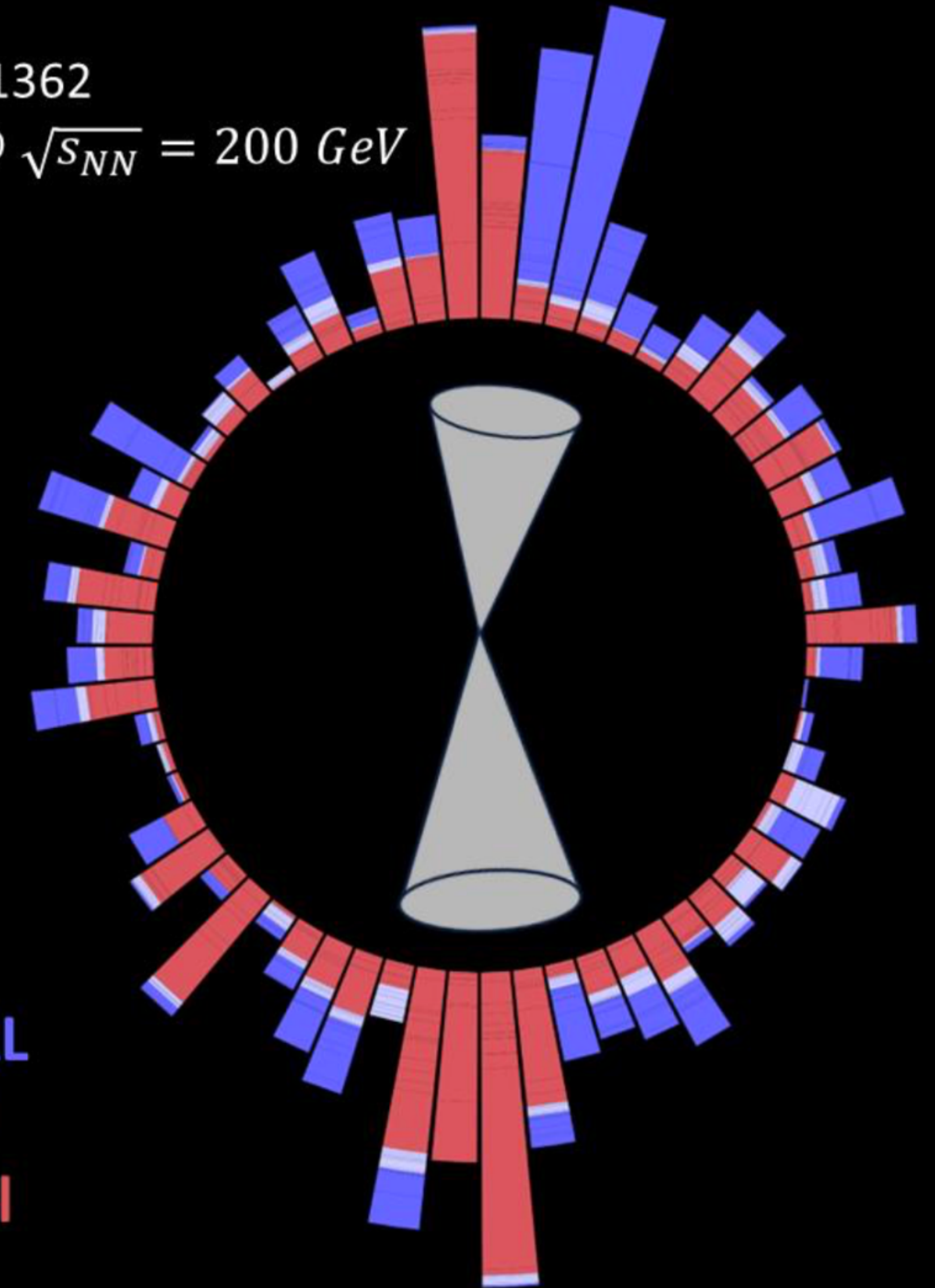
Run/Event: 21615 / 1362

Collisions: Au + Au @ $\sqrt{s_{NN}} = 200 \text{ GeV}$

Peripheral Collision

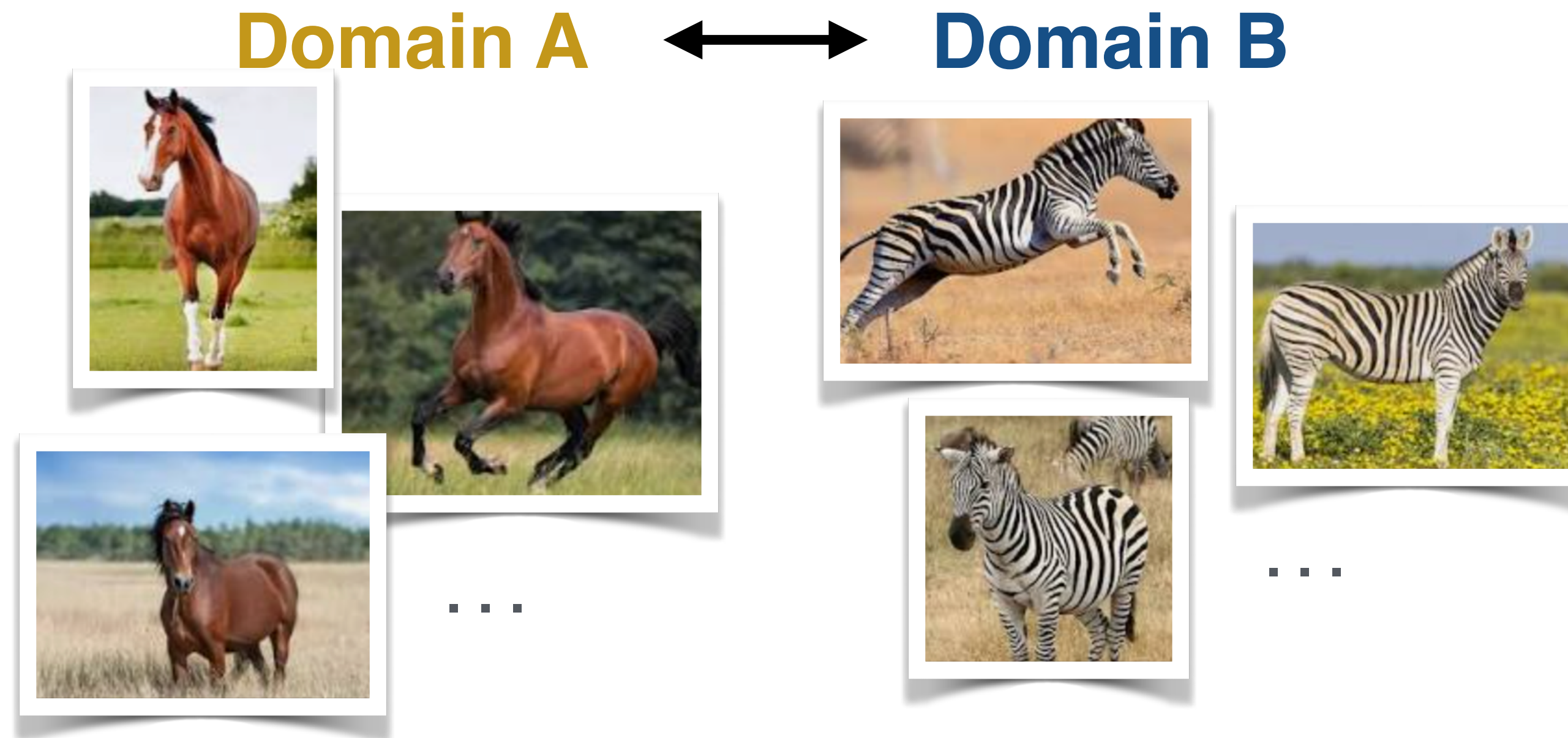
Au+Au collisions
Large background

OHCal
IHCaI
EMCal



Cycle-Consistent Generative Adversarial Network

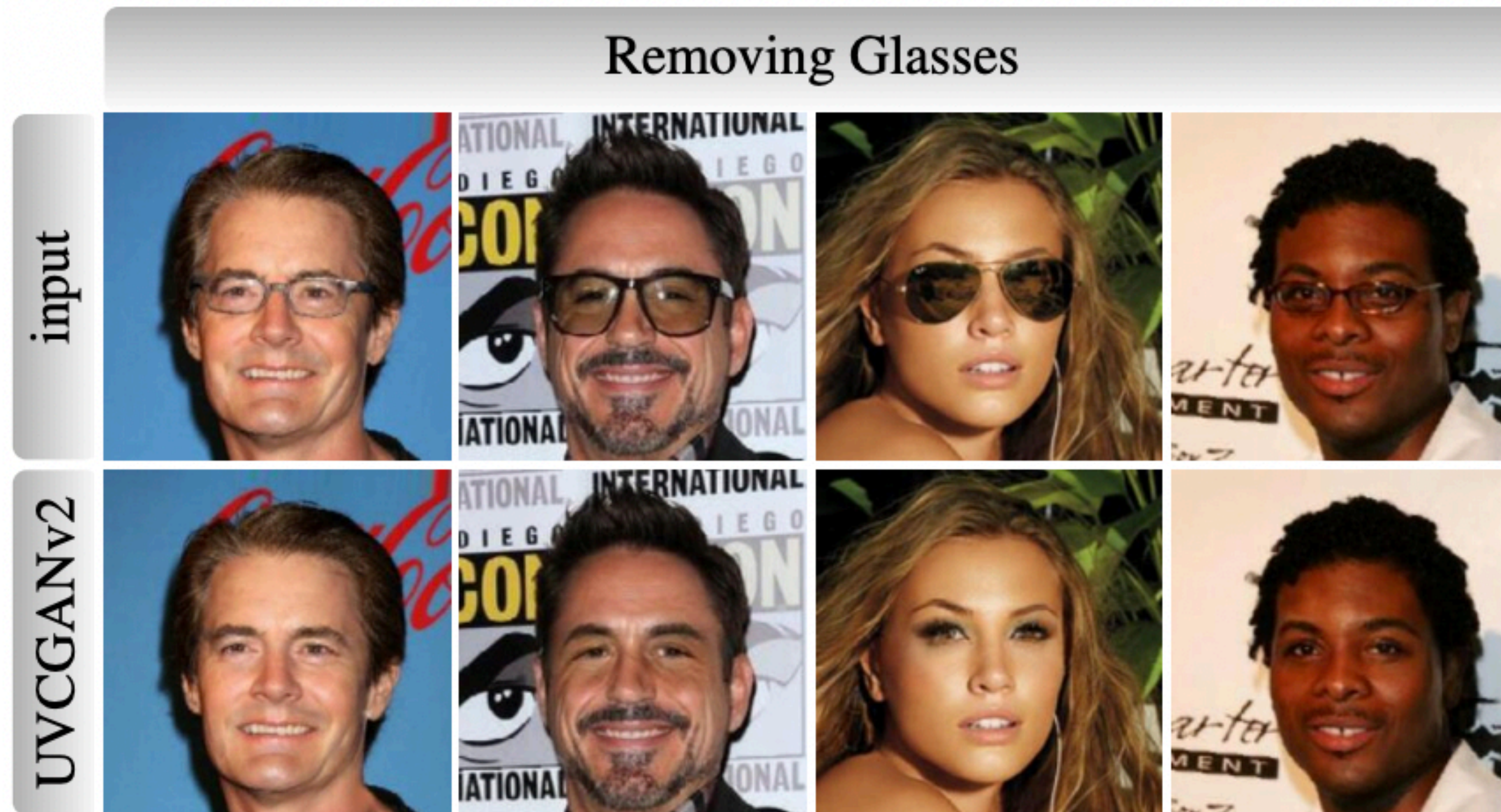
- **Unsupervised** learning, **unpaired** image-to-image translation



arXiv:1703.10593

UVCGAN

- **UVCGAN** (U-net Vision transformer Cycle-consistent Generative Adversarial Network)
 - ➔ **unpaired** image-to-image (I2I) translation; bridging gap between simulation and data reference



[arXiv:2303.16280](https://arxiv.org/abs/2303.16280)

Generative AI for Jet Background Subtraction

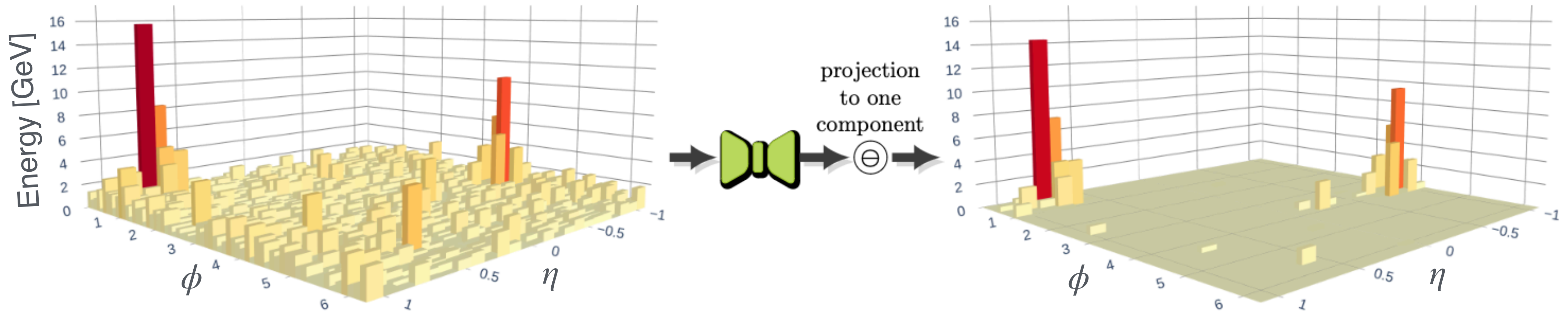
- Jet background subtraction as an “image” using cycle-consistent GAN

- **Unsupervised** generative AI

➔ *preserves the higher dimensional information*

➔ **can use experimental data (without label) for training**
: mitigate model-dependence

↓
Unsupervised learning, **unpaired**
image-to-image translation



[arXiv:2510.23717](https://arxiv.org/abs/2510.23717)

Generative AI for Jet Background Subtraction

- **UVCGAN-S**

- ➔ *unsupervised, unpaired learning allows training directly on data*

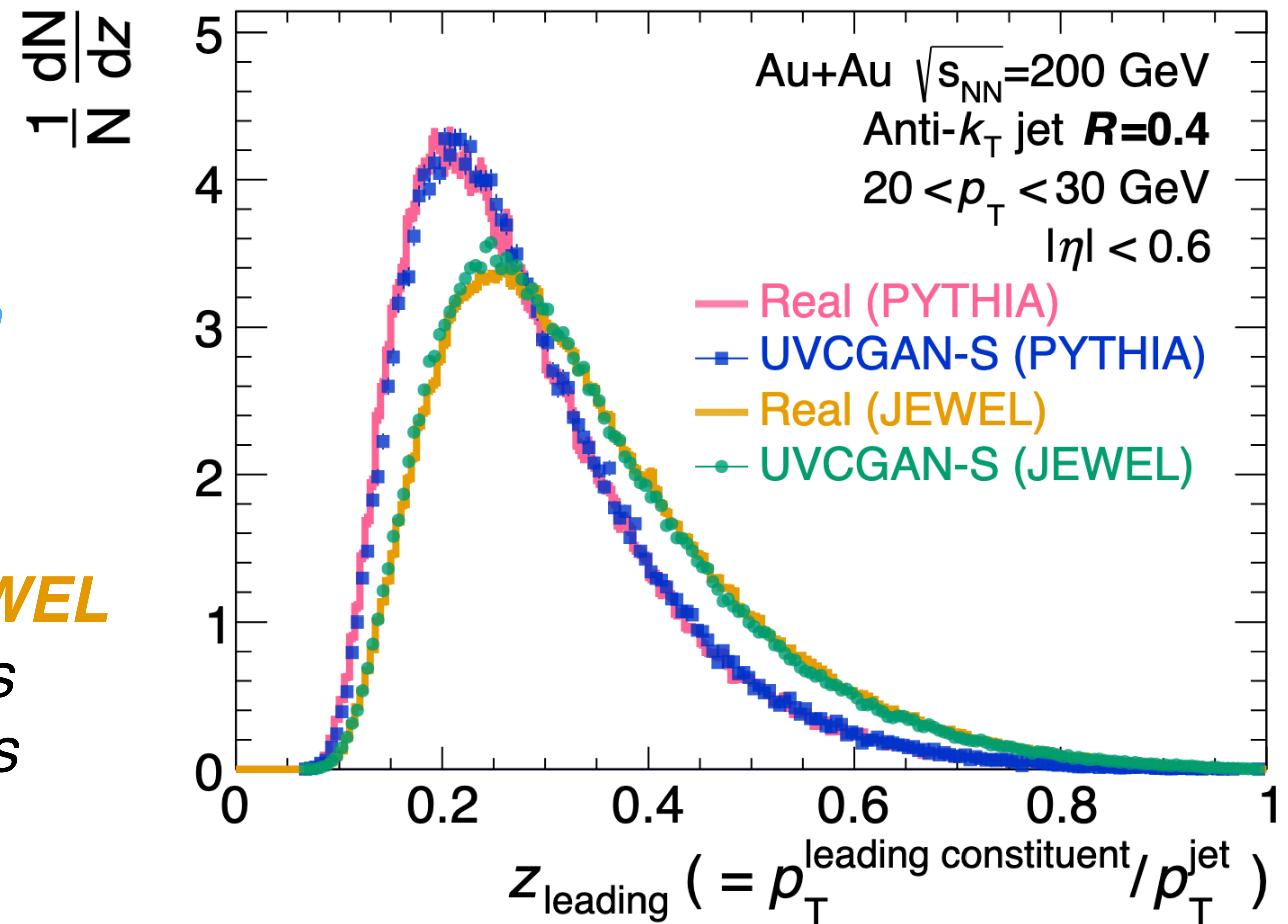
[arXiv:2510.23717](https://arxiv.org/abs/2510.23717)

- ➔ compared to **ICS** or Area methods,
up to ~2x better resolution,
fake rate 80% → 7%,
efficiency 74% → 92%

- ➔ *Robustness to Out-of-Distribution data*

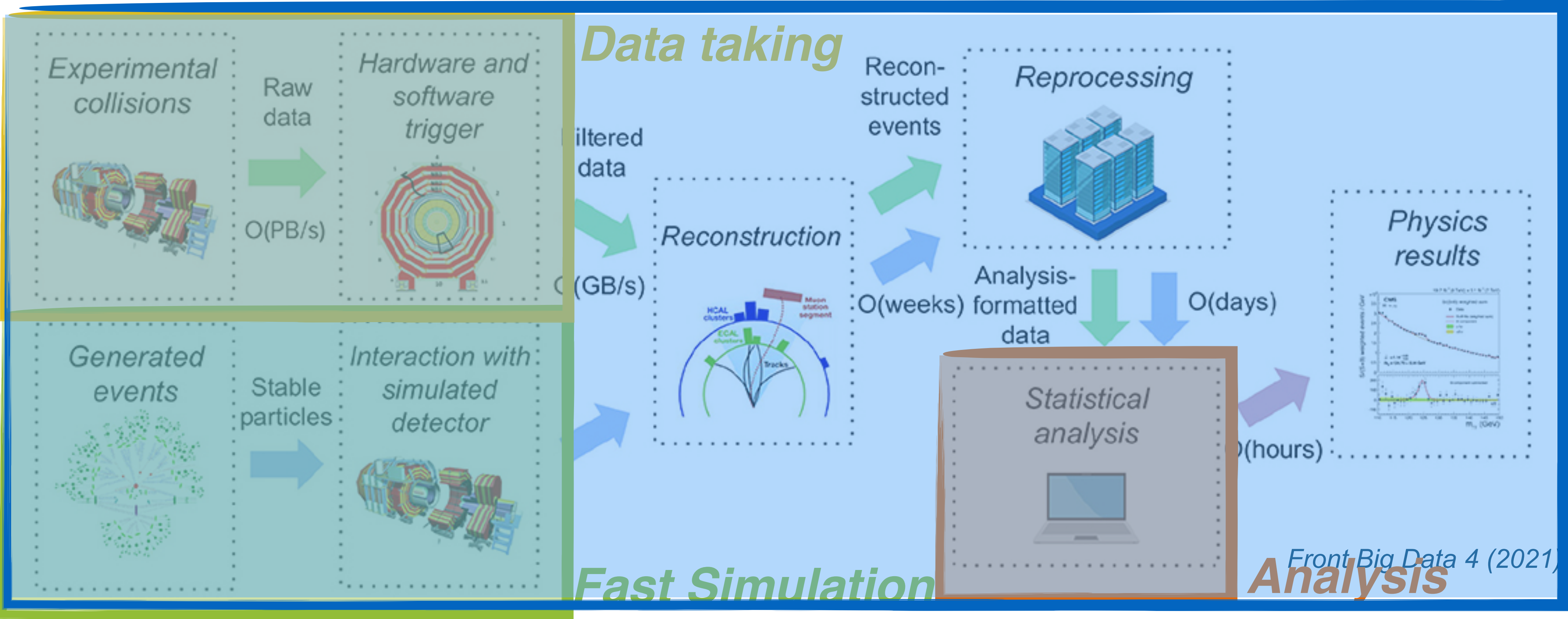
*Trained on **PYTHIA** and tested on **JEWEL***

→ successfully reproduce JEWEL's distributions including substructures



Nuclear and High-Energy Experiment Workflow

Detector design Experiment operation Data analysis Physics discovery



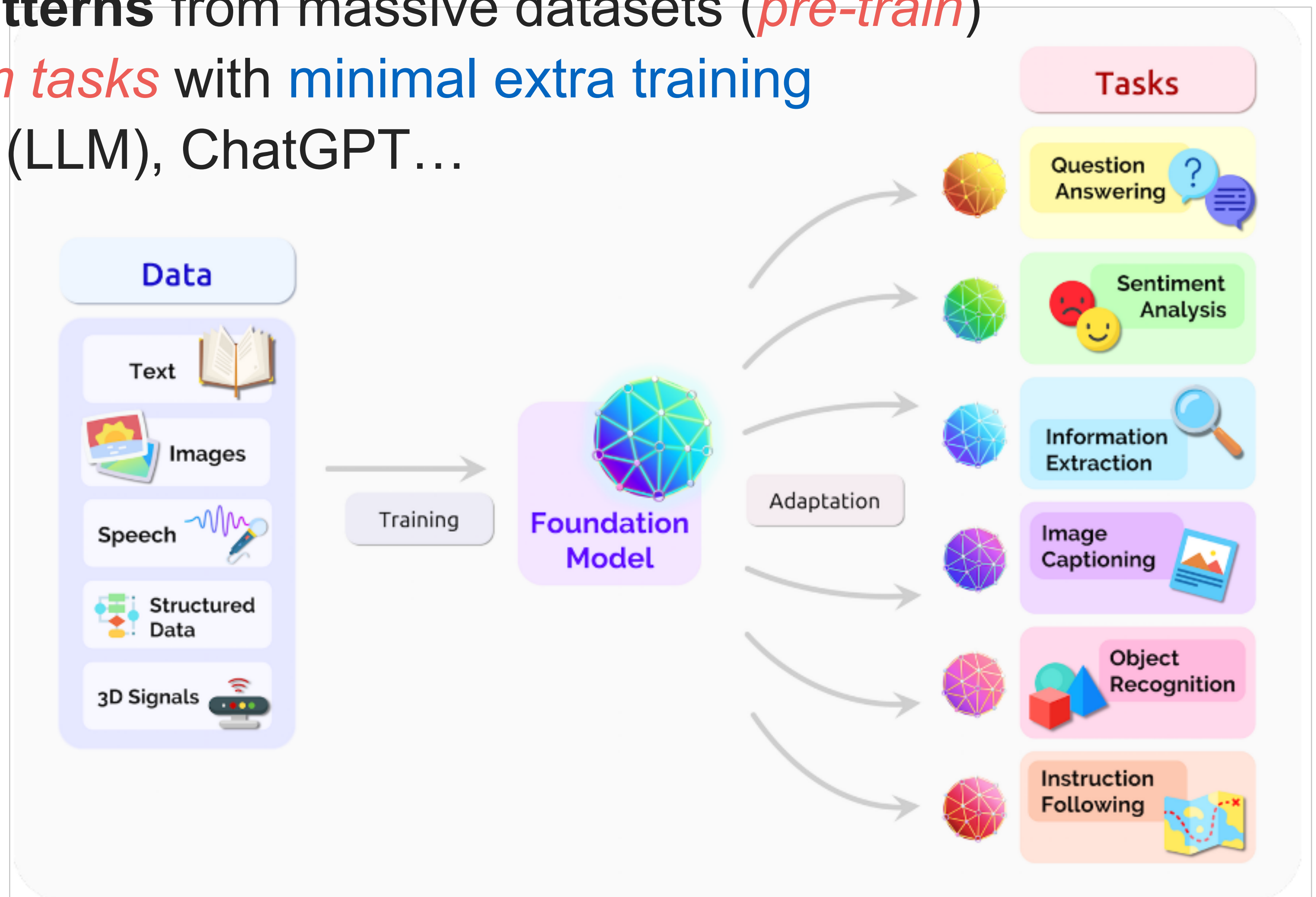
Front Big Data 4 (2021), 661501

Foundation Model !

Foundation Model

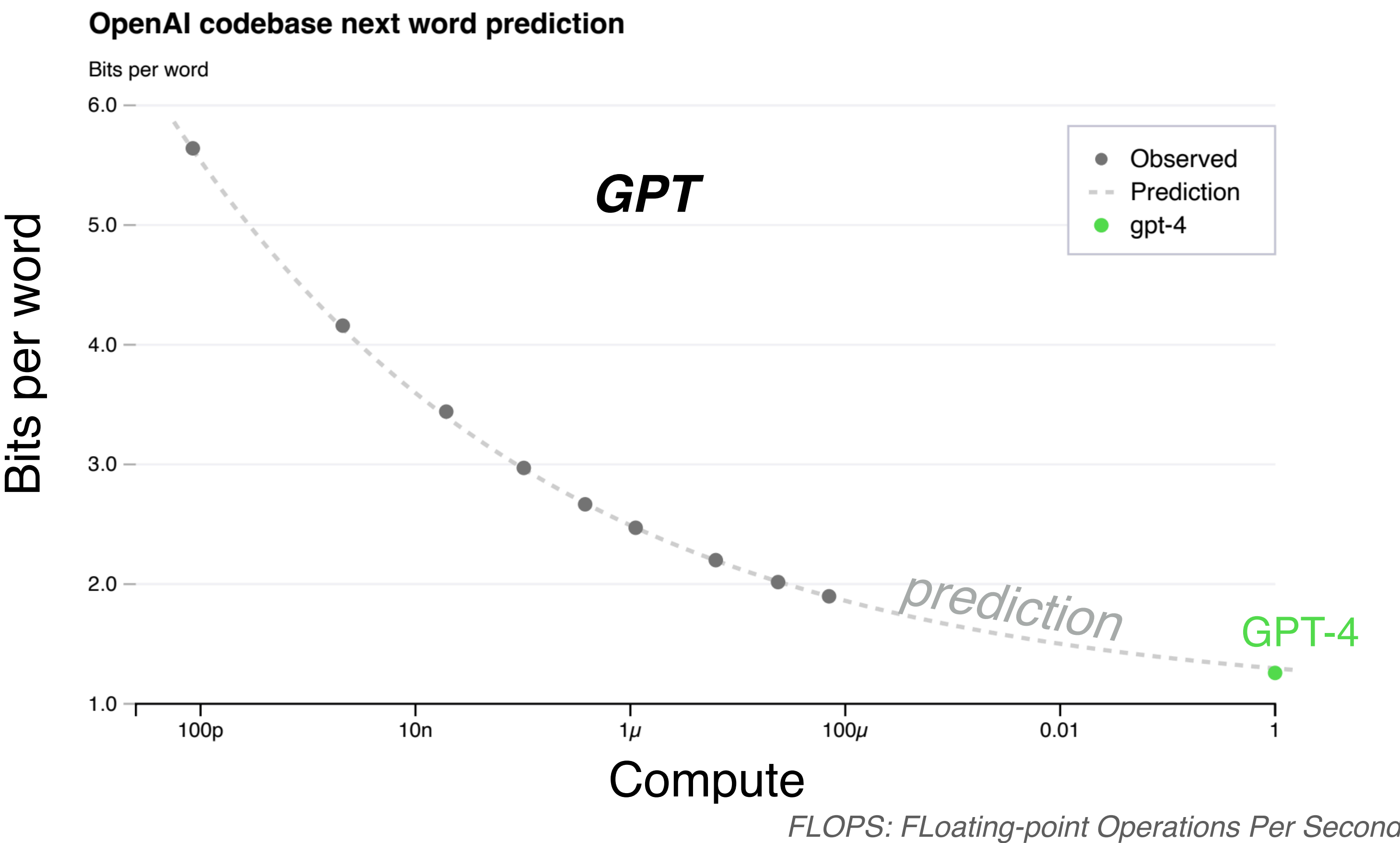
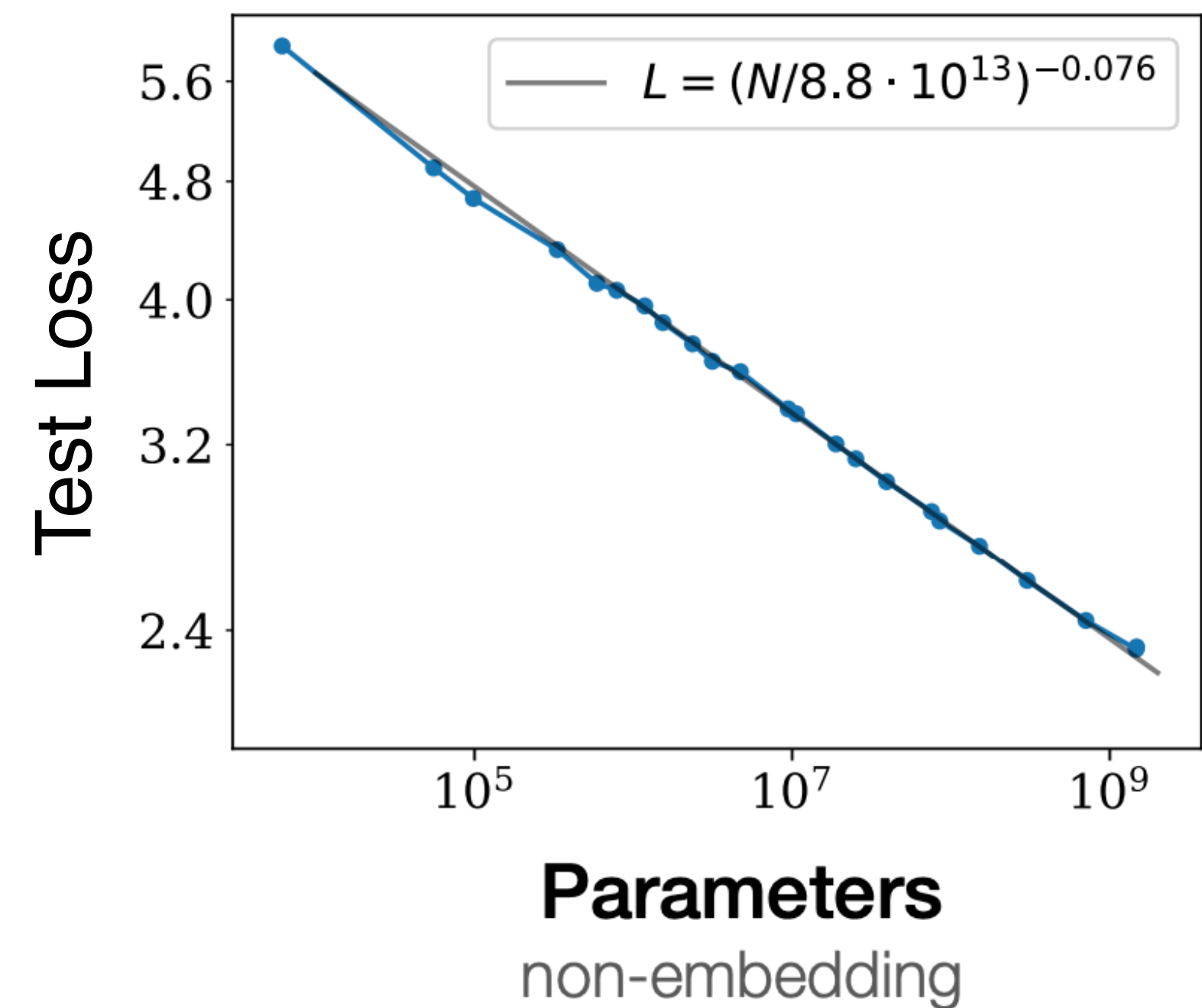
- **Foundation models**

- ➔ learn **broad features and patterns** from massive datasets (*pre-train*)
- ➔ adapted to many *downstream tasks* with **minimal extra training**
- ➔ e.g. Large Language Models (LLM), ChatGPT...



Neural Scaling Laws

- Large language model studies established that performance improves predictably with model size, dataset size, and compute, often following *power law scaling*



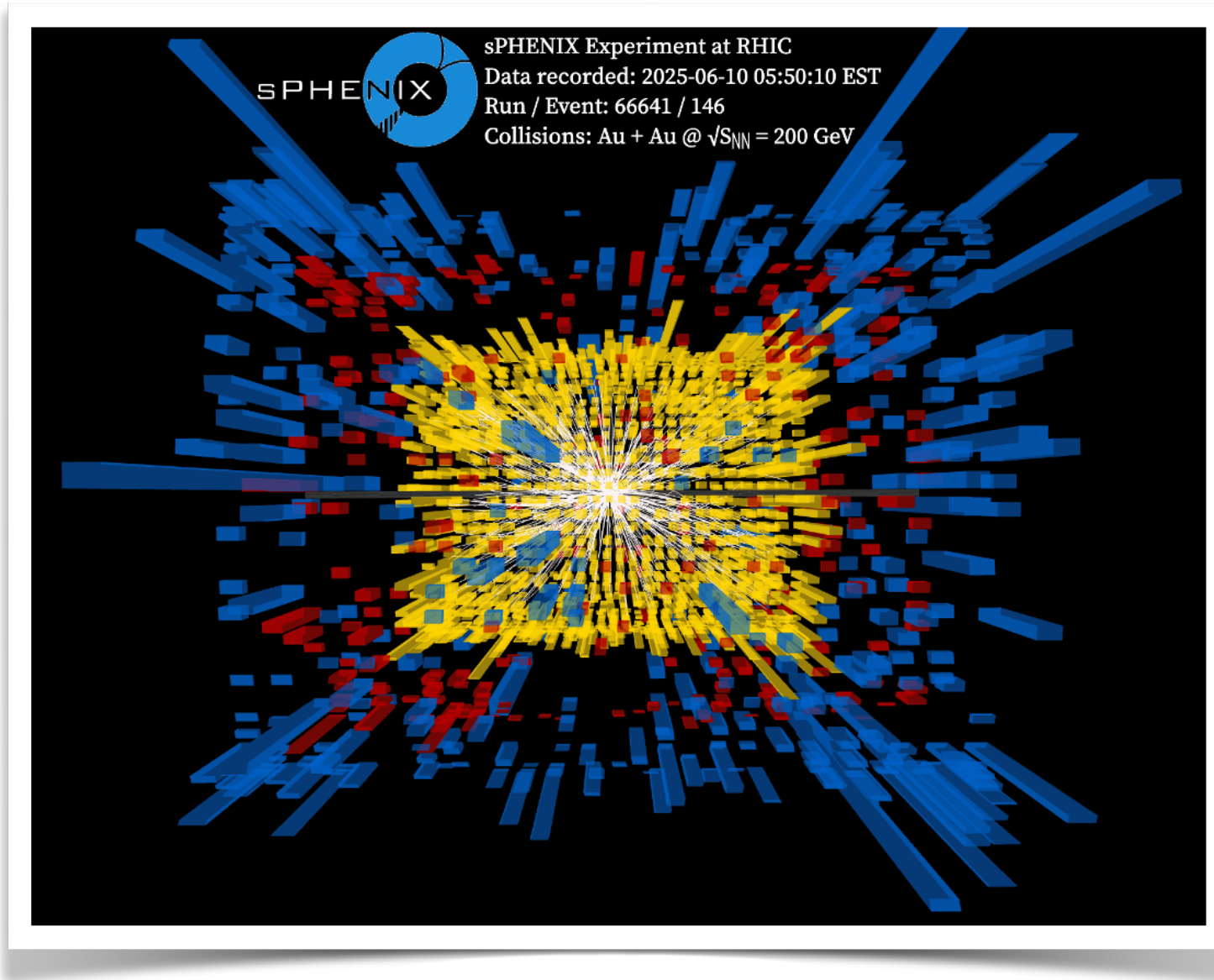
“Scaling Laws for Neural Language Models”
arXiv:2001.08361

Achiam, Josh, et al. "Gpt-4 technical report." arXiv:2303.08774 (2023)

Foundation Model for Nuclear and Particle Physics

Various Downstream Tasks

*Large amount of unlabeled data available
in high energy nuclear and particle physics
e.g. sPHENIX data 300 Petabytes*



Pre-training



FM4
NPP

**minimal
training**



track finding
particle identification
noise removal
global reconstruction
jet tagging
event classification
anomaly detection

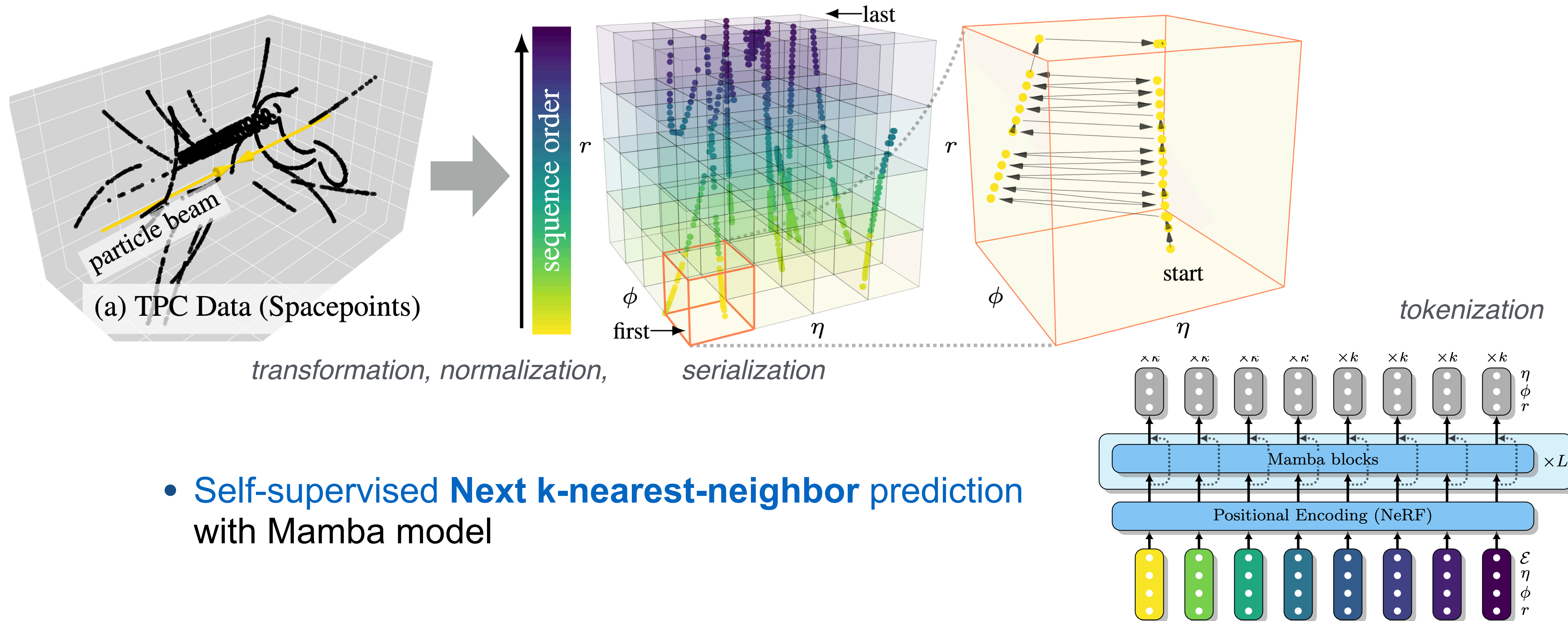
⋮

AI-driven discovery

- **FM4NPP**: a **self-supervised** foundation model with inputs on unlabeled sparse detector-level data
- lightweight **Adapter** for multiple downstream tasks

FM4NPP Training

- Initial study by training with simulated Time Projection Chamber spacepoints (clusters) in 10 million $p+p$ collision events

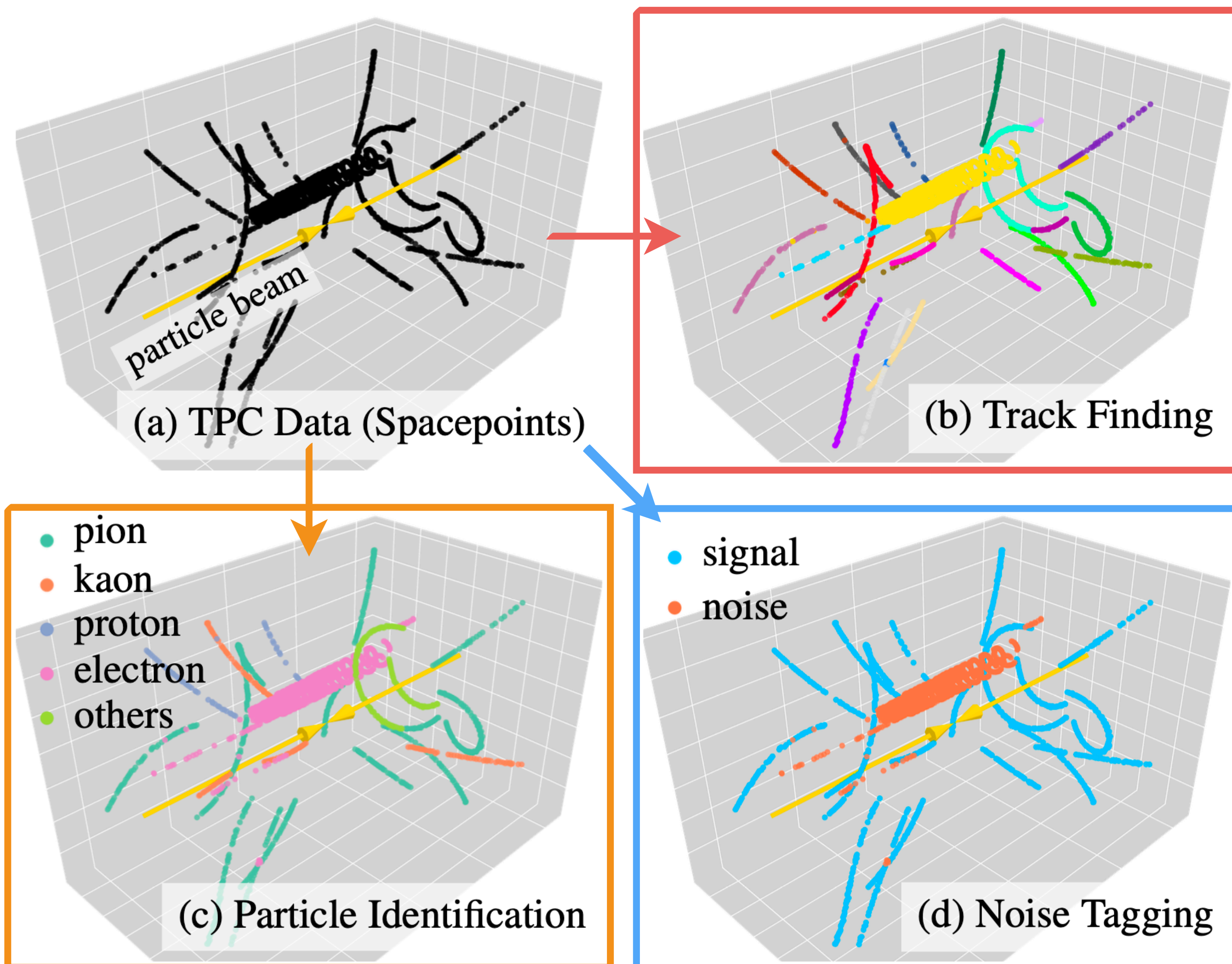


- Self-supervised Next k-nearest-neighbor prediction** with Mamba model

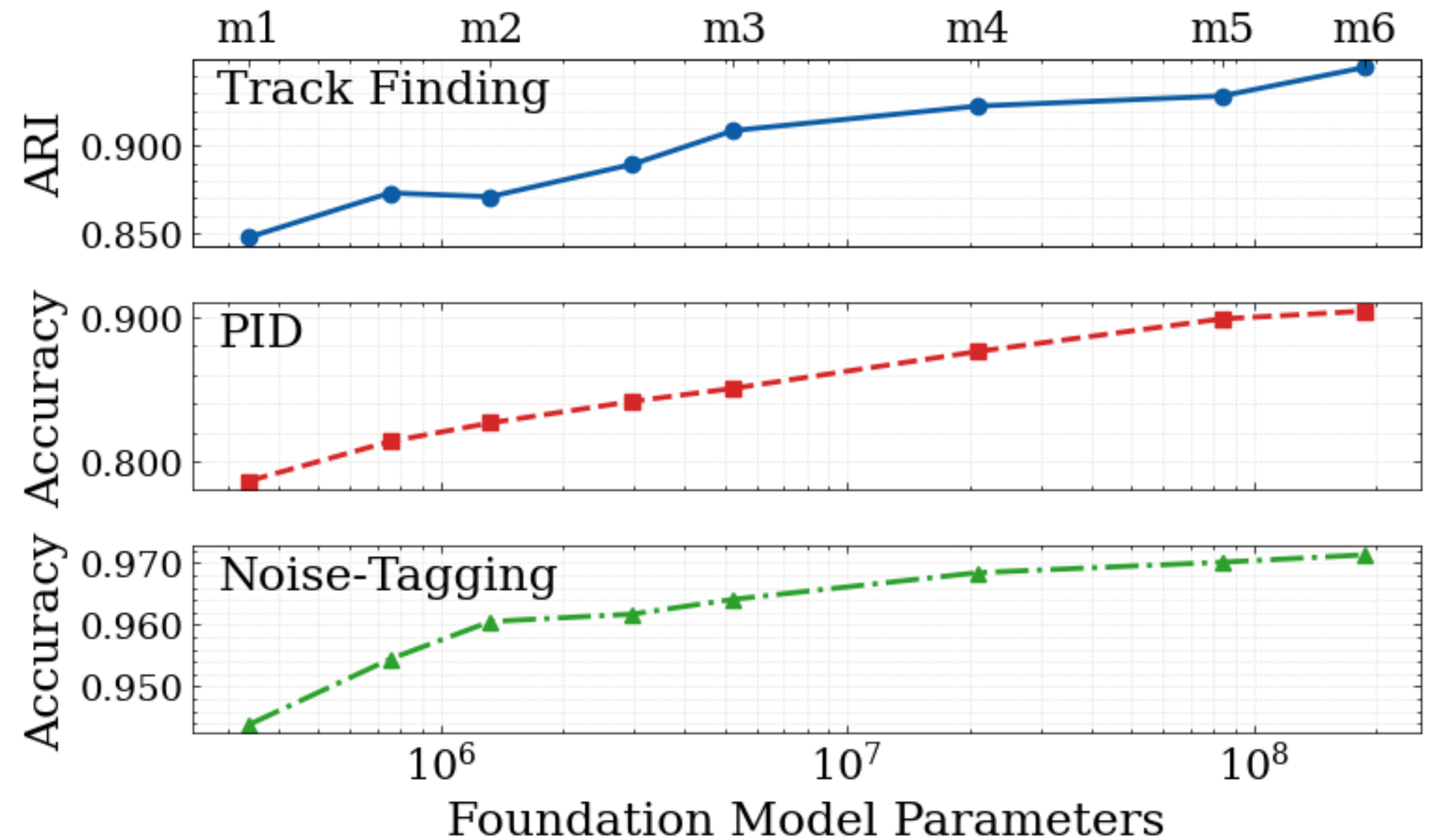
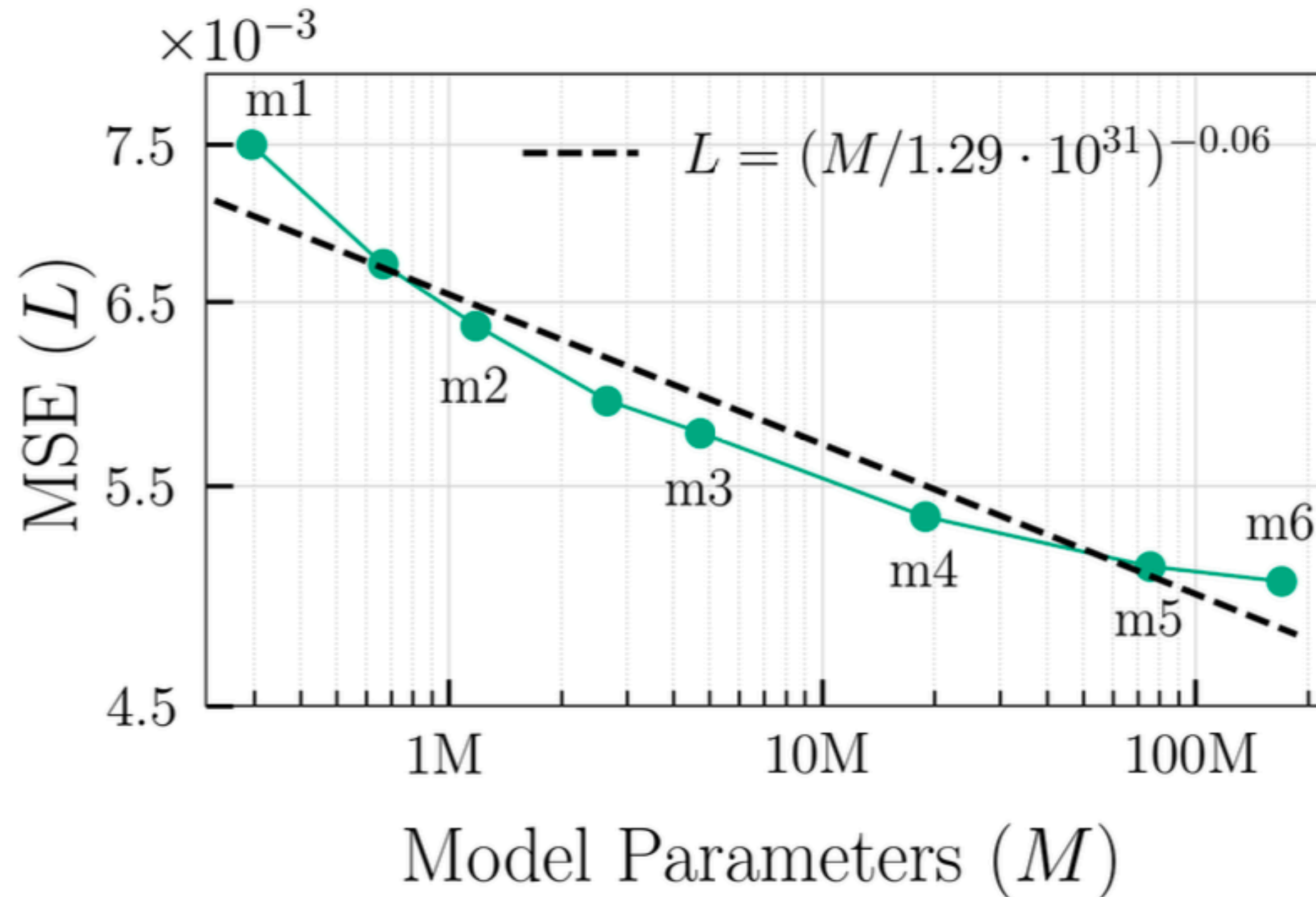
FM4NPP: Downstream Tasks

- The FM backbone is frozen, and only *lightweight decoder* heads are trained
- ➔ **Track Finding**: group spacepoints from the same particle trajectory
- ➔ **Particle Identification**: classify spacepoints based on their particle species
- ➔ **Noise Tagging**: classify spacepoints from low momentum secondary particles (“noise”)

Note: a broader range of downstream tasks will become available as additional detectors are incorporated into the foundation model



FM4NPP: Neural Scaling Behavior



[D. Park, S. Li et al, arXiv:2508.14087](https://arxiv.org/abs/2508.14087)

- Established **self-supervised loss scaling behavior** of **FM4NPP** versus **model parameters**, **data size**, and **compute**
➔ *Consistent with scaling observed in LLMs*

- Demonstrated **scaling behavior for downstream tasks** scaled with foundation model size and data size

FM4NPP: Downstream Task Performance

			Track Finding			
			ARI↑	efficiency↑	purity↑	
model	#trnbl para.					
		EggNet	0.16M	0.7256	74.19%	75.14%
		Exa.TrkX	3.86M	0.8765	91.79%	66.42%
(w/o FM)		AdapterOnly	2.39M	0.7243	78.01%	64.54%
(w/ FM)		FM4NPP (m6)	2.39M	0.9448	96.08%	93.08%

		Particle Identification			Noise Tagging		
model	#trnbl para.	acc.↑	recall↑	pre.↑	acc.↑	recall↑	pre.↑
SAGEConv	0.91M	0.7262	0.4563	0.6502	0.9174	0.7227	0.8165
OneFormer3D	44.95M	0.7701	0.4897	0.5767	0.9646	0.9404	0.8948
..... AdapterOnly	0.74M	0.6631	0.3387	0.6111	0.9111	0.6215	0.8359
..... FM4NPP (m6)	0.74M	0.9039	0.7652	0.8782	0.9713	0.9367	0.9190

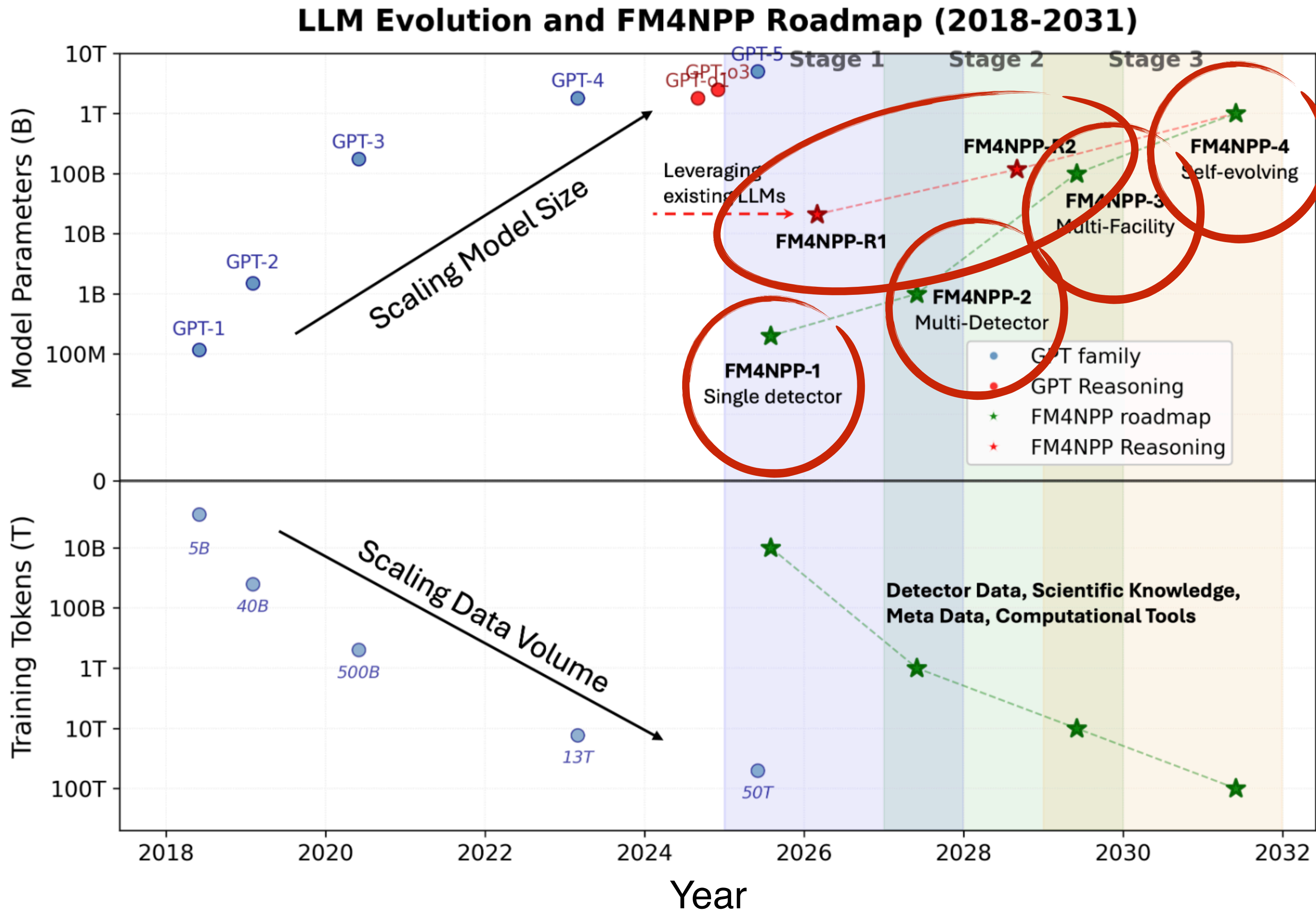
- **FM4NPP** shows clear performance gains compared to an “**AdapterOnly**” baseline
- **Superior performance** of tracking with FM4NPP compared to other state-of-art ML models

[D. Park, S. Li et al, arXiv:2508.14087](#)

FM4NPP: Future Directions

- **Current stage**
 - ➔ initial stage (**FM4NPP-1**) utilized single detector (sPHENIX TPC simulation) with **180M** parameters
 - ➔ Similar scaling behavior, which was first observed in ChatGPT-2
- **Next stages**
 - ➔ Demonstrate on real and more complex data (e.g. heavy ion collisions), 1B parameters (**FM4NPP-1+**)
 - ➔ Multi-detectors (**FM4NPP-2**)
 - ➔ Demonstrate generalizability to multi-facility (**FM4NPP-3**)
 - ➔ Leverage existing LLM reasoning, connect FM4NPP to theory and existing knowledge base (**FM4NPP-R1,2**)
 - ➔ Self-evolving AI (**FM4NPP-4** with an integrated reasoning model)

[Roadmap Whitepaper \[ssrn.5467666\]](#)



Conclusion / Take-home Message

- AI and ML are **opening new directions** and **driving innovation** across many areas of physics
- **AI can enhance the entire scientific workflow**, from detector operation and data processing to simulation, analysis, and scientific discovery
- **Foundation models** offer a promising path toward broadly reusable tools for nuclear and particle physics, potentially enabling an impact analogous to that of large language models
- Realizing this potential requires **physics-informed models, reliable uncertainty quantification, rigorous validation**, and **close collaboration between physicists and AI experts**



JET
(HARD PROBE)

ISOLATED
PHOTON
(HARD PROBE)

JET
(HARD PROBE)

HUMAN INTELLIGENCE
+ MACHINE INTELLIGENCE

DISCOVERING THE FUNDAMENTAL
BUILDING BLOCKS OF NATURE



UNITED BY CURIOSITY
DRIVEN BY DISCOVERY



Thank you for listening

BACKUP

Agentic AI

- AI has the potential to **increase scientific productivity** across the full workflow
- This includes **coding, analysis design, analysis execution, validation, and scientific writing**

Resummation of the C-Parameter Sudakov Shoulder Using Effective Field Theory

Matthew D. Schwartz^{1,2}

¹*Department of Physics, Harvard University, Cambridge, MA 02138, USA*

²*Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)*

schwartz@g.harvard.edu

AI RESEARCH ASSISTANT: Claude Opus 4.5 (Anthropic)

January 7, 2026

Abstract

The C-parameter distribution in e^+e^- annihilation exhibits a kinematic shoulder at $C = 3/4$, where three-parton final states reach their maximum and a fourth parton is required to exceed it. This boundary generates large logarithms that must be resummed. Using soft-collinear effective theory, we derive a factorization theorem involving new jet and soft functions specific to the C-parameter measurement, in which soft radiation contributes quadratically in transverse momentum. This quadratic structure explains the step discontinuity at leading order. We compute all ingredients at one loop, validate against Monte Carlo, and present matched NLL+NLO results. Unlike thrust and heavy jet mass, the C-parameter has no Sudakov–Landau pole, making momentum-space resummation straightforward. All calculations, numerical analysis, and manuscript preparation were performed by Claude, an AI assistant developed by Anthropic, working under physicist supervision.

[M. Schwartz arXiv:2601.02484](#)

Agentic AI – Physicist Collaboration in Experimental Particle Physics: A Proof-of-Concept Measurement with LEP Open Data

Anthony Badea,^a Yi Chen,^b Marcello Maggi,^c Yen-Jie Lee,^d Electron-Positron Alliance

^a*University of Chicago, Enrico Fermi Institute, 60637 IL, USA*

^b*Vanderbilt University, Nashville, Tennessee, USA*

^c*Istituto Nazionale di Fisica Nucleare, Bari Division, BA 70126, Italy*

^d*Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

E-mail: anthony.badea@cern.ch

ABSTRACT: We present an AI agentic measurement of the thrust distribution in e^+e^- collisions at $\sqrt{s} = 91.2$ GeV using archived ALEPH data. The analysis and all note writing is carried out entirely by AI agents (OpenAI Codex and Anthropic Claude) under expert physicist direction. A fully corrected spectrum is obtained via Iterative Bayesian Unfolding and Monte Carlo based corrections. This work represents a step toward a theory–experiment loop in which AI agents assist with experimental measurements and theoretical calculations, and synthesize insights by comparing the results, thereby accelerating the cycle that drives discovery in fundamental physics. Our work suggests that precision physics, leveraging the open LEP data and advanced theoretical landscape, provides an ideal testing ground for developing advanced AI systems for scientific applications.

[A. Badea, Y. Chen et al, arXiv:2603.05735](#)

Automating High Energy Physics Data Analysis with LLM-Powered Agents

Eli Gendreau-Distler,^{1,2,*} Joshua Ho,^{1,2,†} Dongwon Kim,^{1,2,‡} Luc
Tomas Le Pottier,^{1,2,§} Haichen Wang,^{1,2,¶} and Chengxi Yang^{1,2,**}

¹*Department of Physics, University of California, Berkeley, Berkeley, CA 94720, USA*

²*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

We present a proof-of-principle study demonstrating the use of large language model (LLM) agents to automate a representative high energy physics (HEP) analysis. Using the Higgs boson diphoton cross-section measurement as a case study with ATLAS Open Data, we design a hybrid system that combines an LLM-based supervisor–coder agent with the **Snakemake** workflow manager. In this architecture, the workflow manager enforces reproducibility and determinism, while the agent autonomously generates, executes, and iteratively corrects analysis code in response to user instructions. We define quantitative evaluation metrics – success rate, error distribution, costs per specific task, and average number of API call for the task – to assess agent performance across multi-stage workflows. To characterize variability across architectures, we benchmark a representative selection of state-of-the-art LLMs, spanning the *Gemini* and *GPT-5* series, the *Claude* family, and leading open-weight models. While the workflow manager ensures deterministic execution of all analysis steps, the final outputs still show stochastic variation. Although we set the temperature to zero, other sampling parameters (e.g., top-p, top-k) remained at their defaults, and some reasoning-oriented models internally adjust these settings. Consequently, the models do not produce fully deterministic results. This study establishes the first LLM-agent-driven automated data-analysis framework in HEP, enabling systematic benchmarking of model capabilities, stability, and limitations in real-world scientific computing environments. The baseline code used in this work is available at <https://huggingface.co/HWresearch/LLM4HEP>. This work was accepted as a poster at the Machine Learning and the Physical Sciences (ML4PS) workshop of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025). The initial submission was made on August 30, 2025.

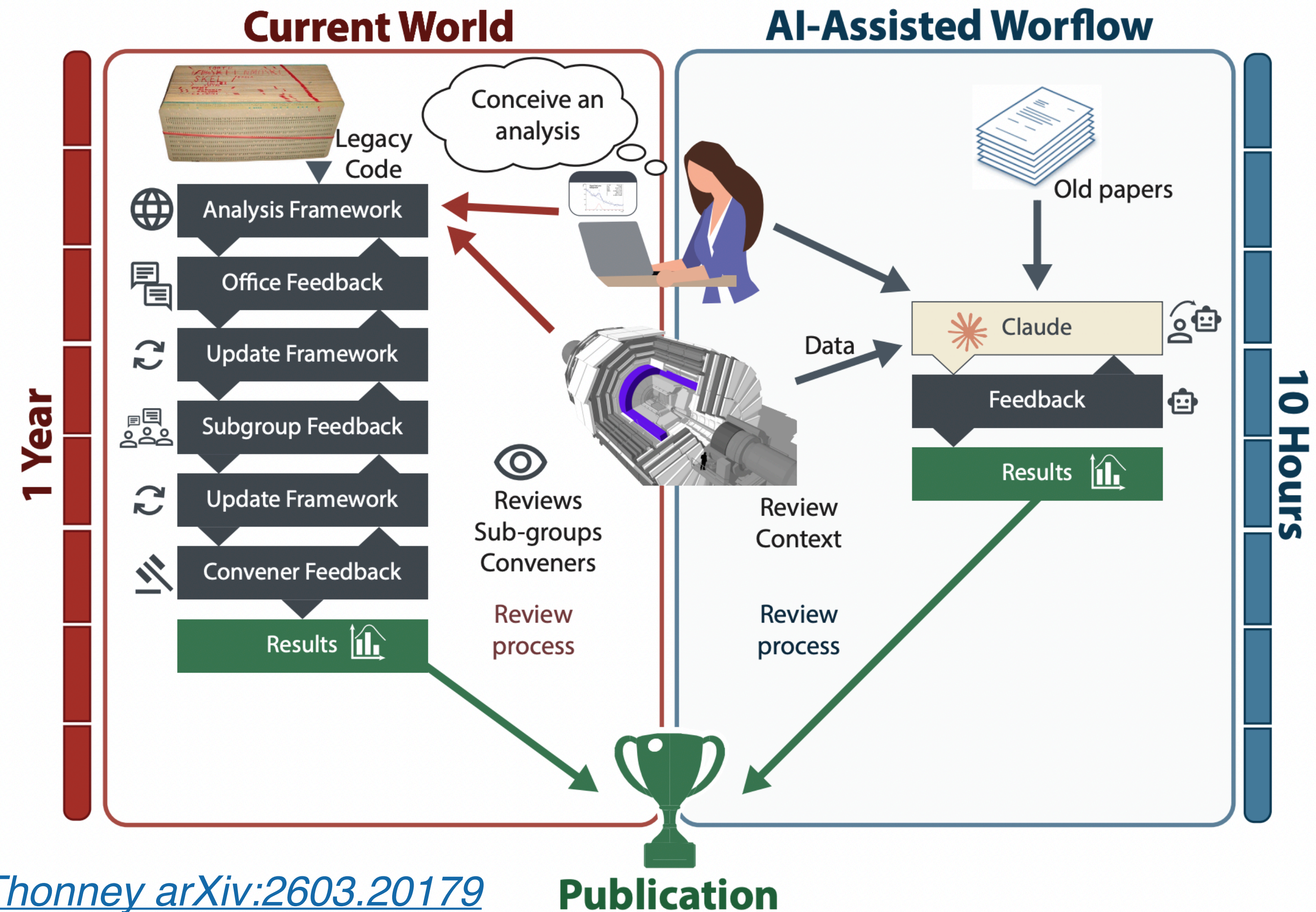
[E. Gendreau-Distler et al, arXiv:2512.07785](#)

Agentic AI

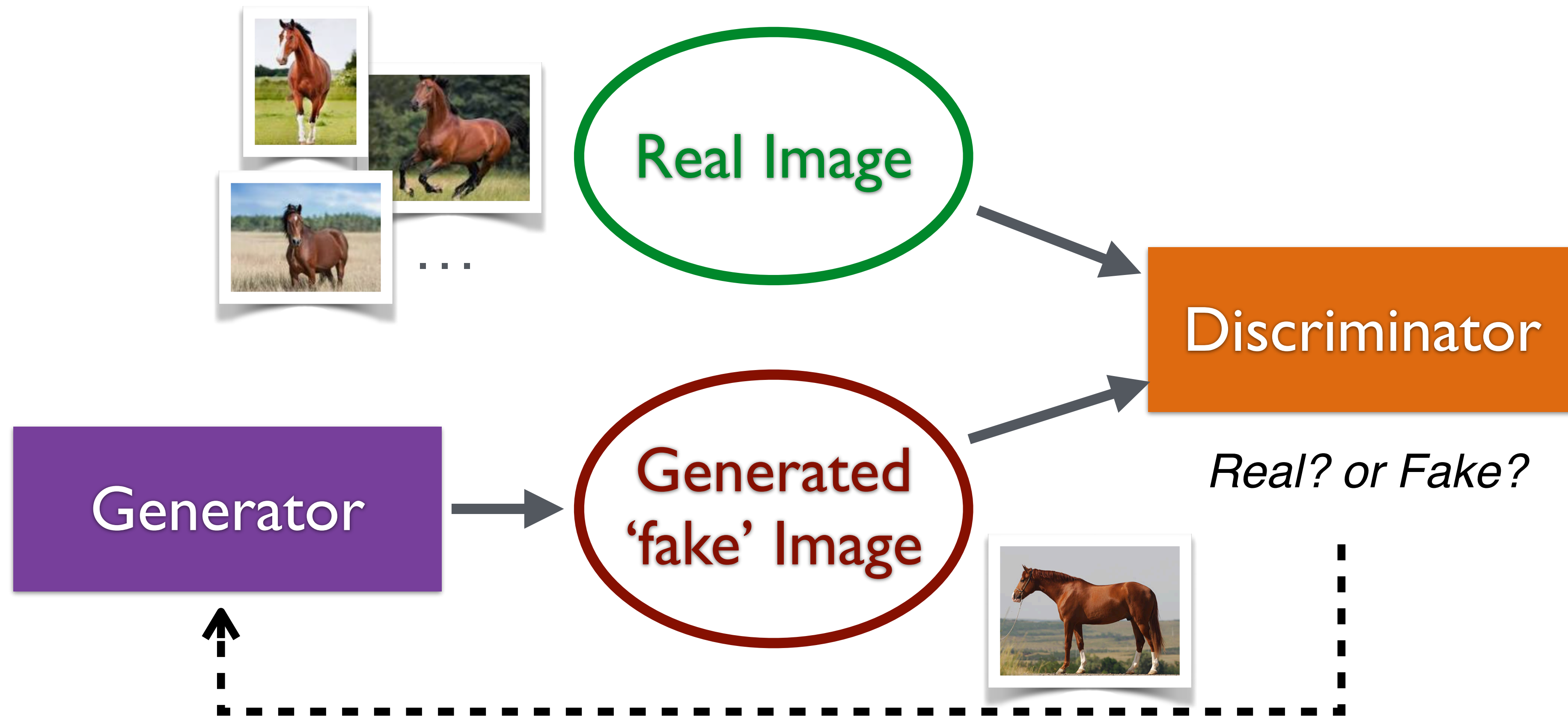
- AI has the potential to **increase scientific productivity** across the full workflow
- This includes **coding, analysis design, analysis execution, validation, and scientific writing**

- **Key consideration:**

- ➔ How do we identify and control biases from AI models?
- ➔ How do we reliably maintain scientific quality?
- ➔ How do we review and assess a rapidly growing volume of AI-assisted work?
- ➔ ...



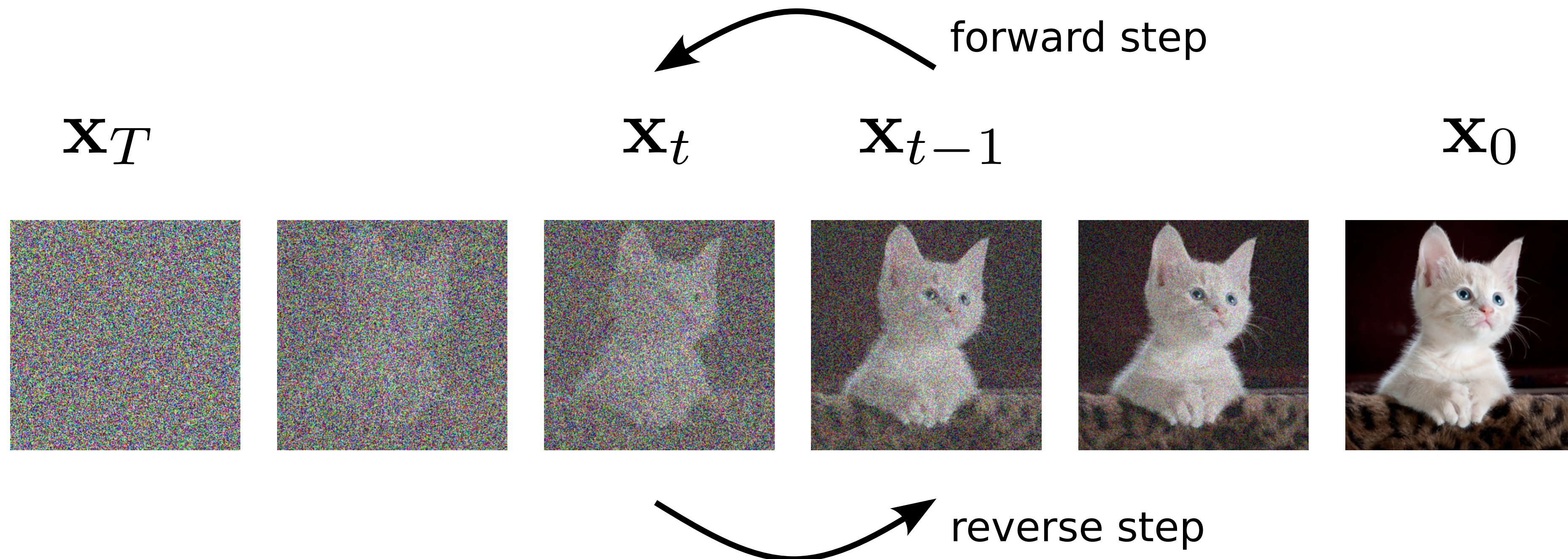
Generative Adversarial Network (GAN)



- Two neural networks (**Generator** and **Discriminator**) *compete* against each other
 - ➔ **feedback** to each other to improve
 - ➔ the process is repeated until **Discriminator** can no longer distinguish fake generated by **Generator** from real

Denoising Diffusion Probabilistic Model (DDPM)

- **Forward** process: gradually adds **Gaussian noise** to the data
- **Reverse** process: a neural network learns to reverse this process; start from **pure noise**, **gradually remove noise** using the learned denoising model; **generate data** similar to the **original images**
- In real application, $O(1,000)$ steps are used for high quality

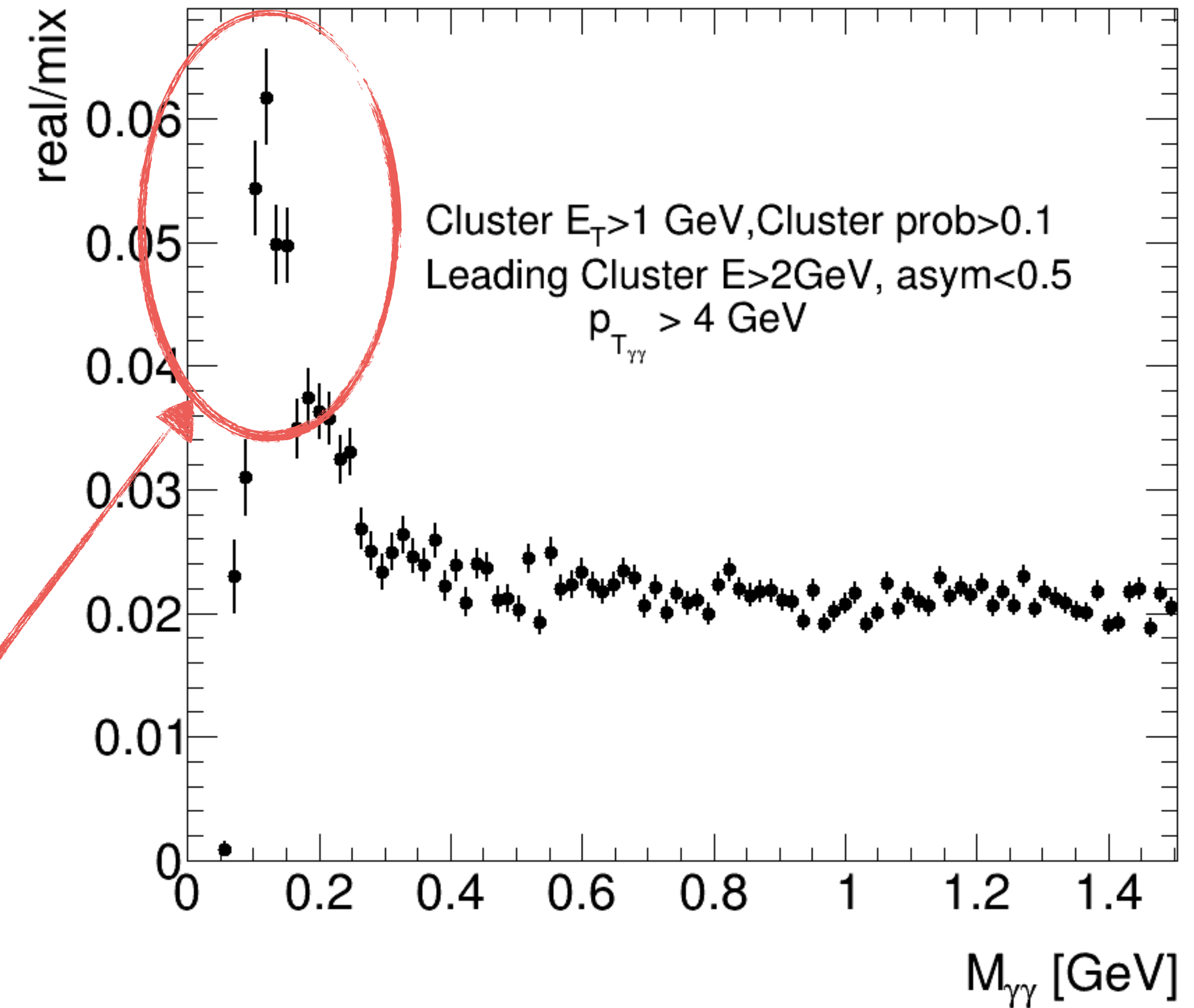


Whole-Event and Full-Calorimeter Simulation using DDPM

- Can DDPM describe more *complex features* of heavy ion collisions?

➔ **Resonance, collective motion (flow)**
can be reproduced by **DDPM (!)**

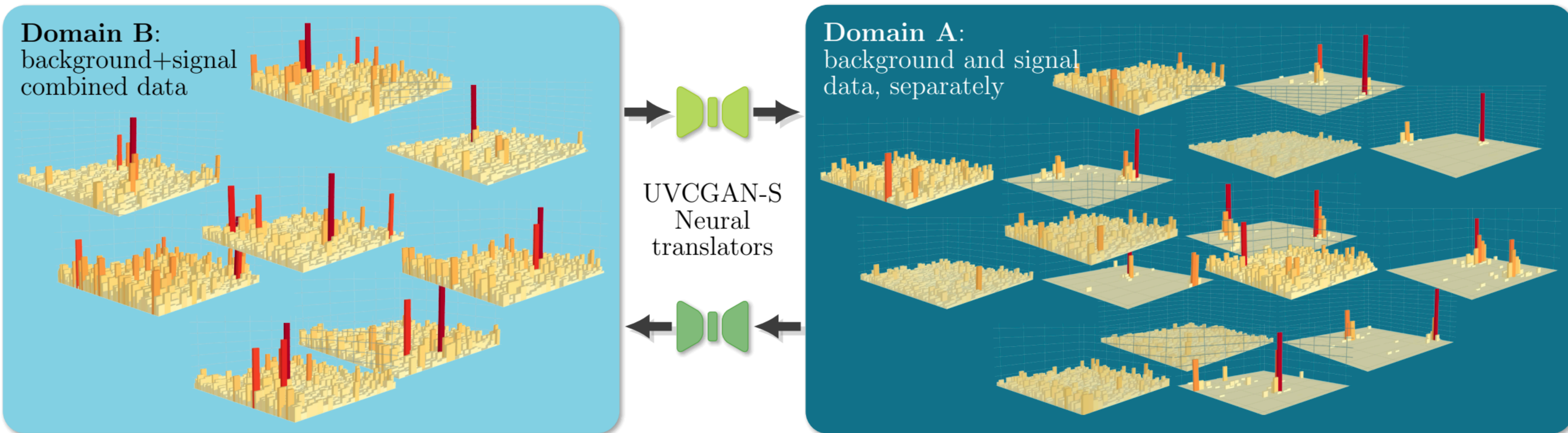
π^0 peak reconstructed using
simulation samples generated by DDPM



UVCGAN-S: Unpaired Unsupervised Training

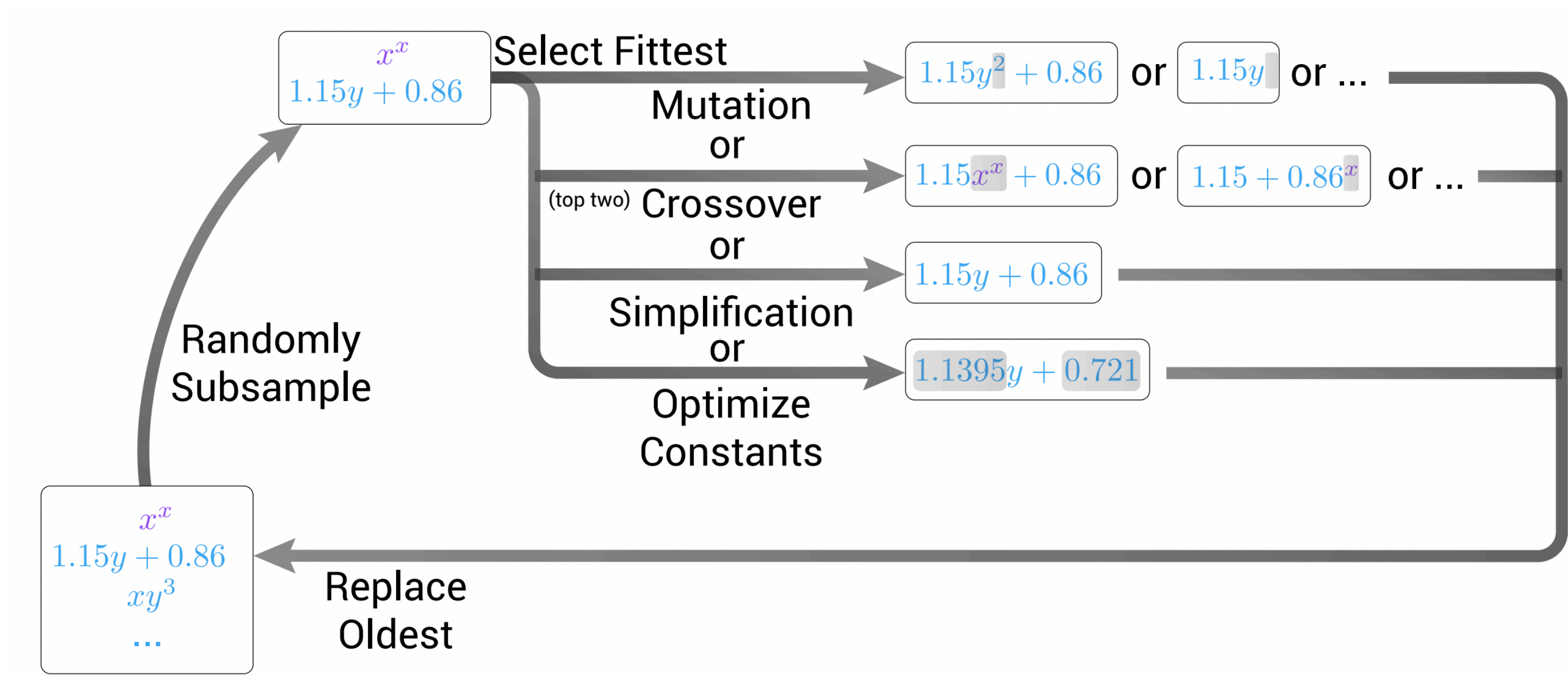
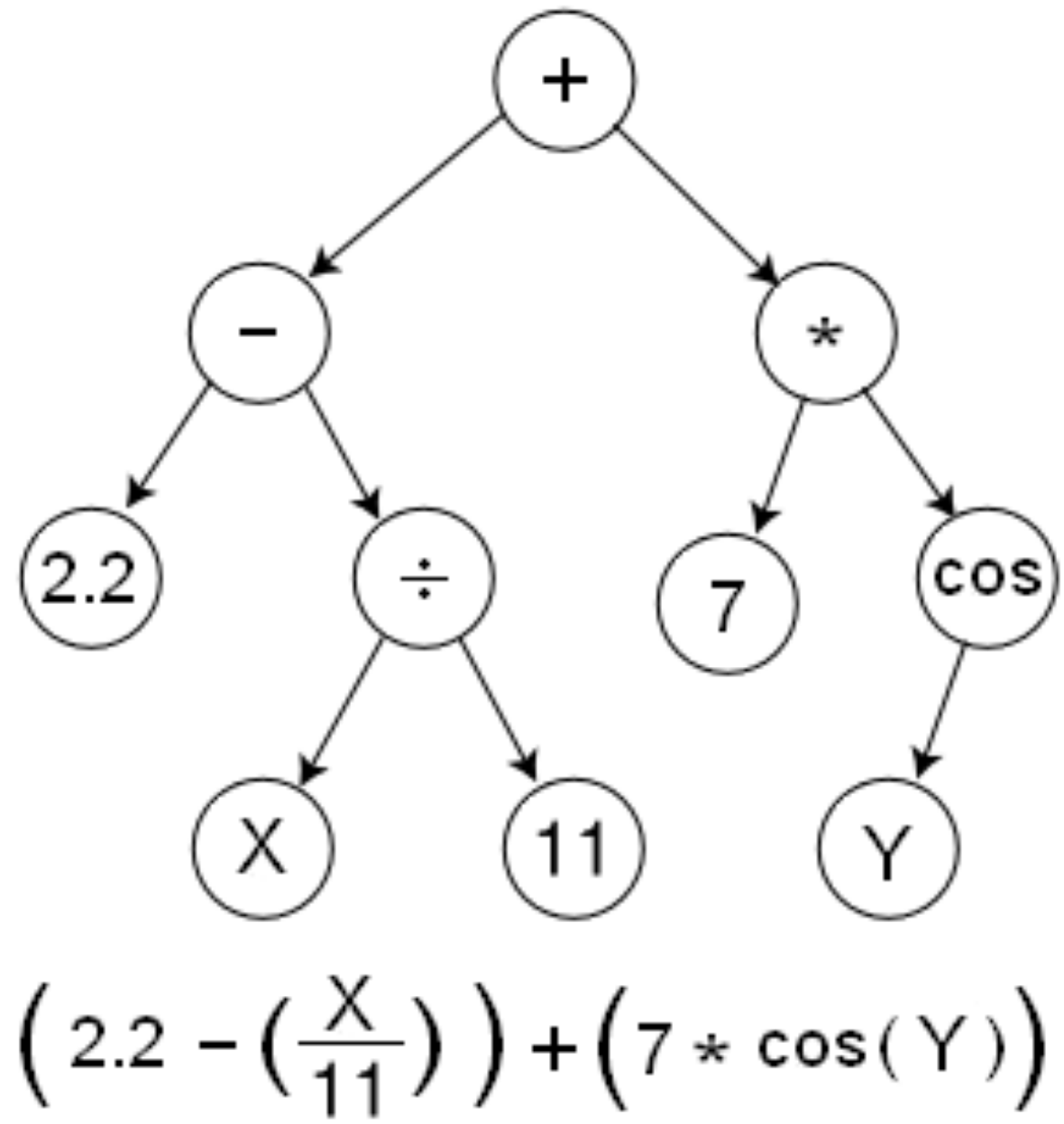
- As a proof-of-principle, UVCGAN-S model was trained on images of *EMCal + HCAL combined* calorimeter energies in (η, ϕ) with three samples
 - **PYTHIA**: Inclusive Jet $\hat{p}_T = 30$ GeV
 - **HIJING**: 0-10% Centrality HIJING
 - **PYTHIA+HIJING**: sum of calorimeter energies of PYTHIA and HIJING

[arXiv:2510.23717](https://arxiv.org/abs/2510.23717)



Symbolic Regression

- Symbolic Regression finds analytic expressions
- Use DGLAP evolution as a controlled benchmark
- Use **physics-informed loss** functions with Mellin-space evolution
- produces interpretable analytic formulas, not black-box predictions



FM4NPP Training

- Initial study by training with simulated Time Projection Chamber spacepoints (clusters) in 10 million $p+p$ collision events
- **Hierarchical Raster Scan (HRS)** serializes spacepoints to 1D sequence
- **Mamba** model: recent State Space Model (SSM) that improves computational efficiency while maintaining long-range sequence modeling capabilities
- **Self-supervised Next k-nearest-neighbor prediction**: instead of ordinary autoregression or “next token” prediction

