The 2nd "AI+HEP in East Asia" Workshop

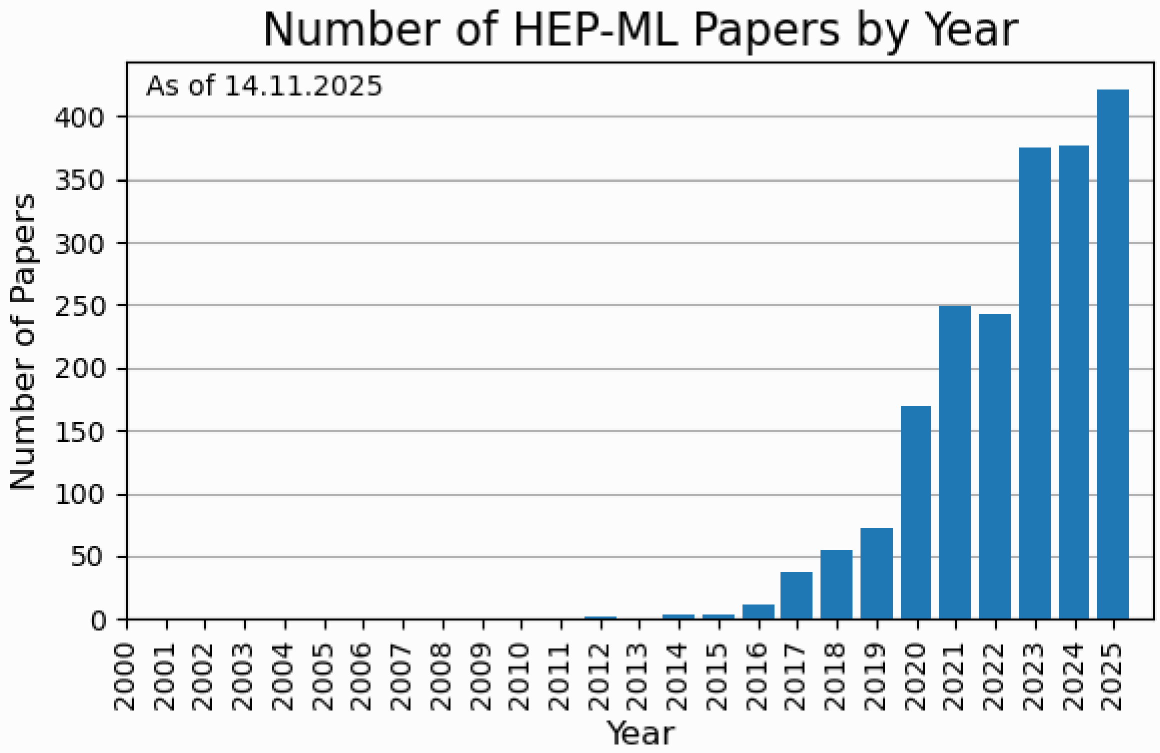
AI and Machine Learning Application in Experimental High Energy Physics

Liang Li

Shanghai Jiao Tong University

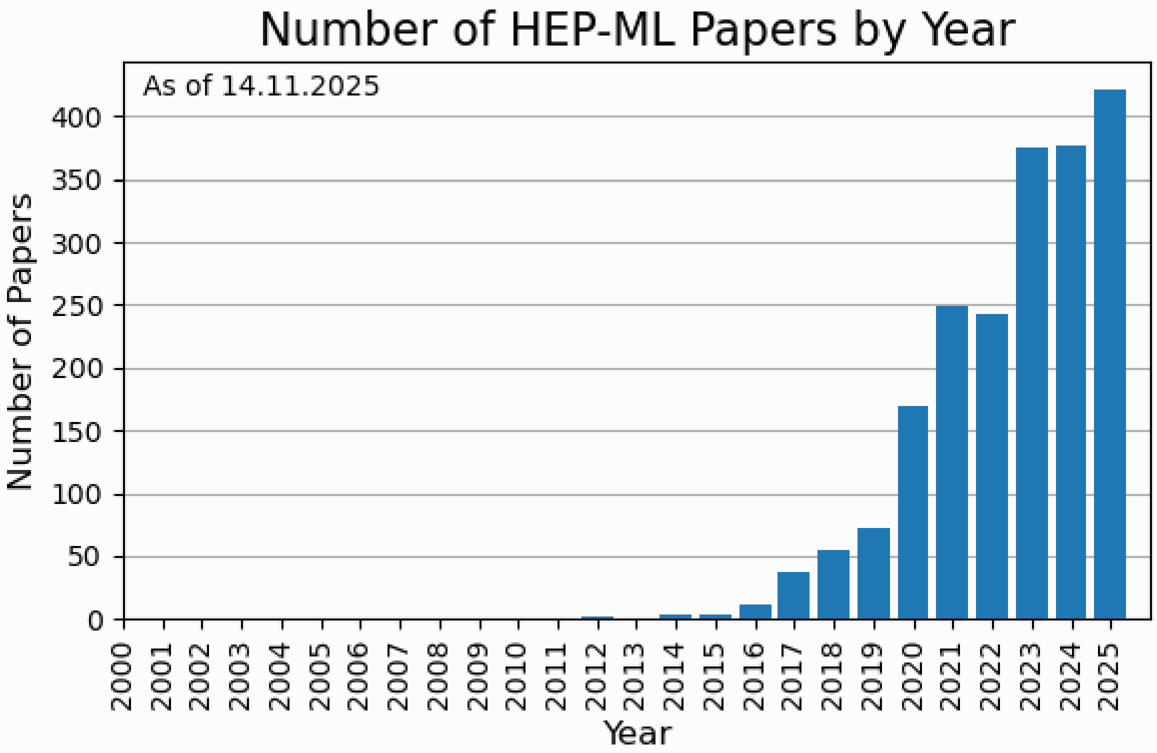
AI and ML in High Energy Physics

AI and ML in High Energy Physics



[Living Review of Machine Learning for Particle Physics]

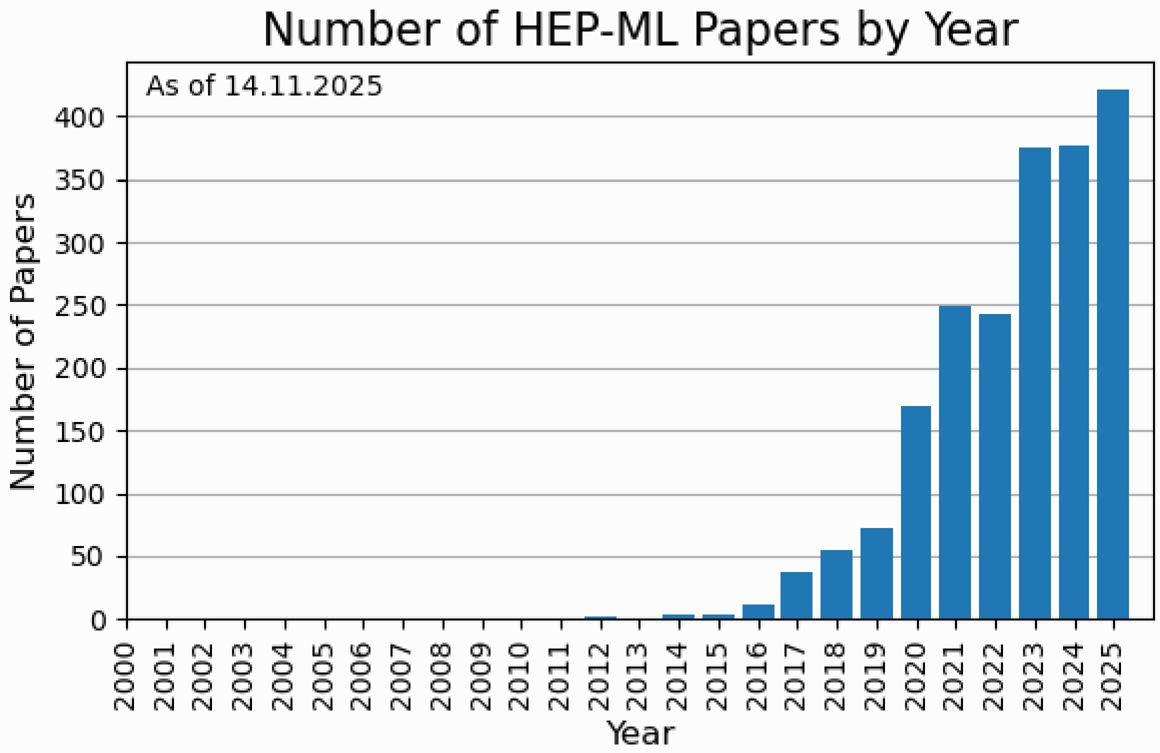
AI and ML in High Energy Physics



[Living Review of Machine Learning for Particle Physics]

- **Do you know how many papers linked on the landing page?**

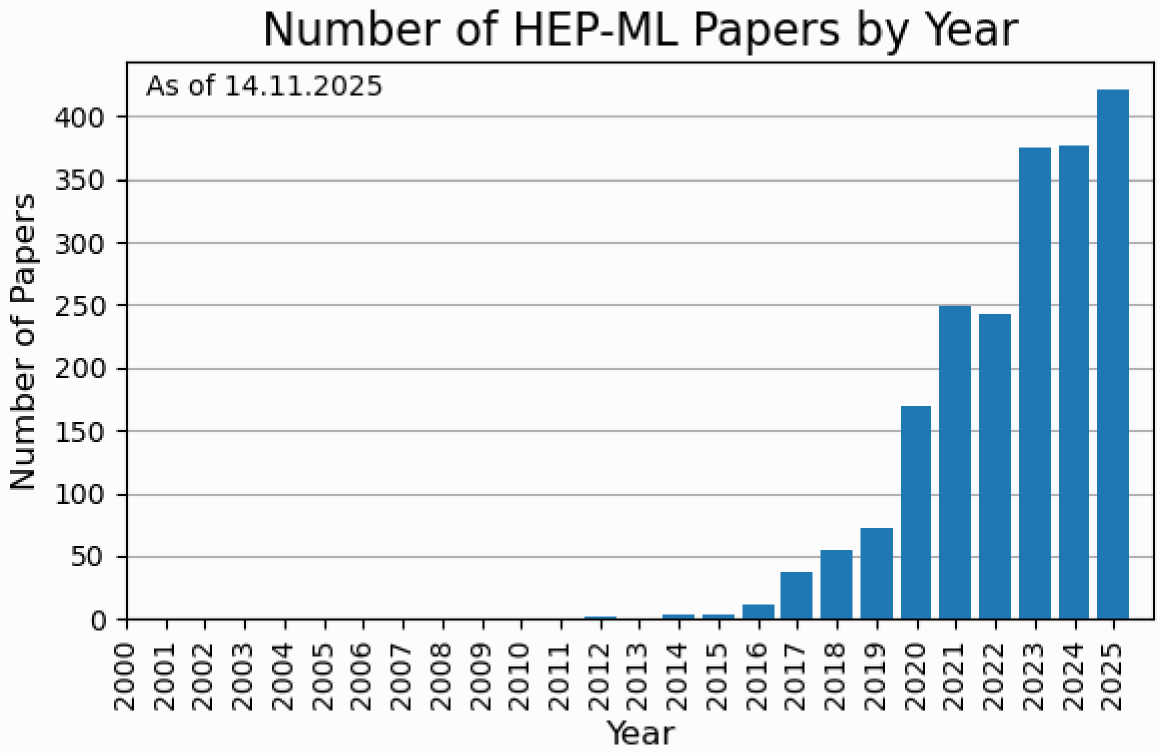
AI and ML in High Energy Physics



[Living Review of Machine Learning for Particle Physics]

- **Do you know how many papers linked on the landing page?**
- **Read them all, or, let AI do it for you ☺**

AI and ML in High Energy Physics



Living Review of Machine Learning for Particle Physics

- Do you know how many papers linked on the landing page?
- Read them all, or, let AI do it for you ☺

Impossible to cover everything in one talk

- **Highly selective and apologize for missing many important work**

✓Classifier (Supervised)

✓Classifier (Supervised)

✓Self-guided Detection/Search (Weakly Supervised/Unsupervised)

✓Classifier (Supervised)

✓Self-guided Detection/Search (Weakly Supervised/Unsupervised)

✓Reconstruction

AI and ML in Experimental High Energy Physics

SJTUPA 上海交通大学物理与天文学院

✓**Classifier (Supervised)**

✓**Self-guided Detection/Search (Weakly Supervised/Unsupervised)**

✓**Reconstruction**

✓**Simulation**

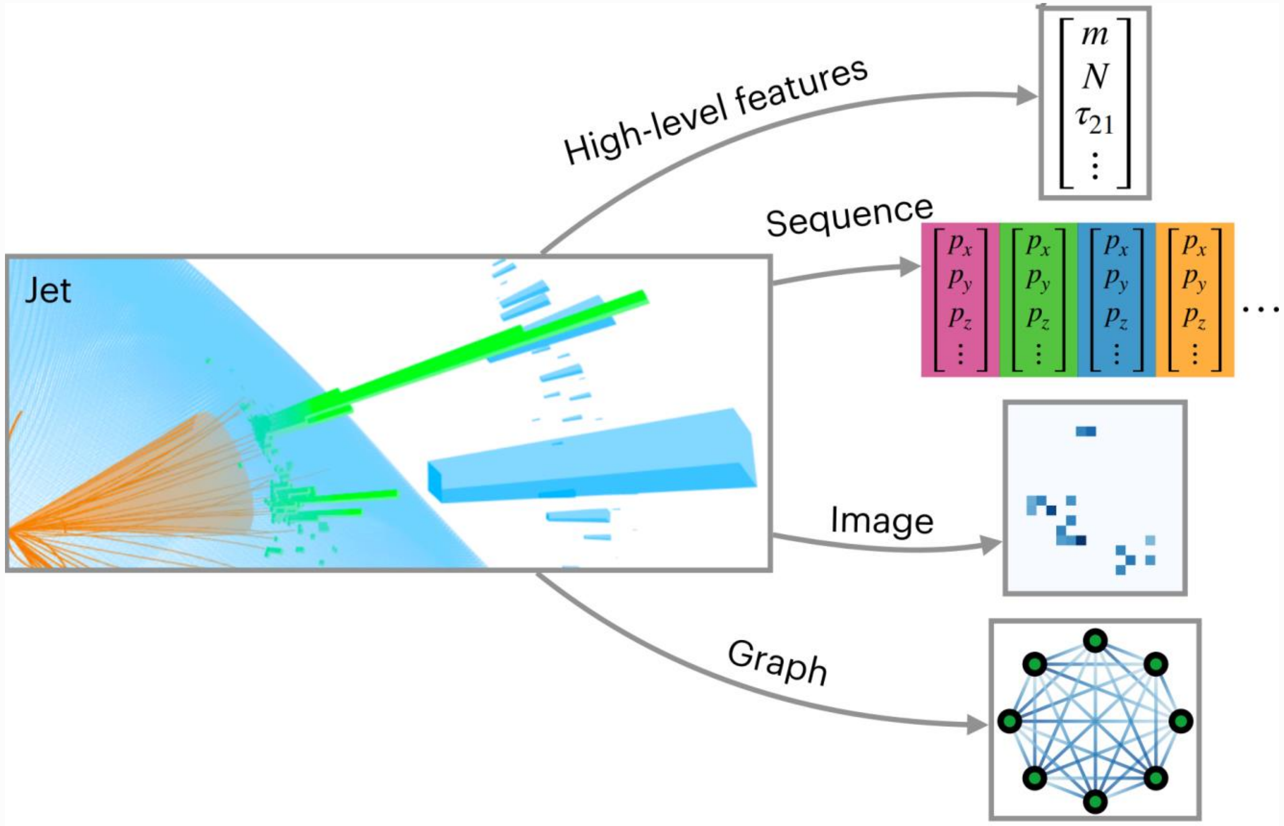
- ✓Classifier (Supervised)
- ✓Self-guided Detection/Search (Weakly Supervised/Unsupervised)
- ✓Reconstruction
- ✓Simulation
- ✓Language Model

SJTUPA 上海交通大学物理与天文学院

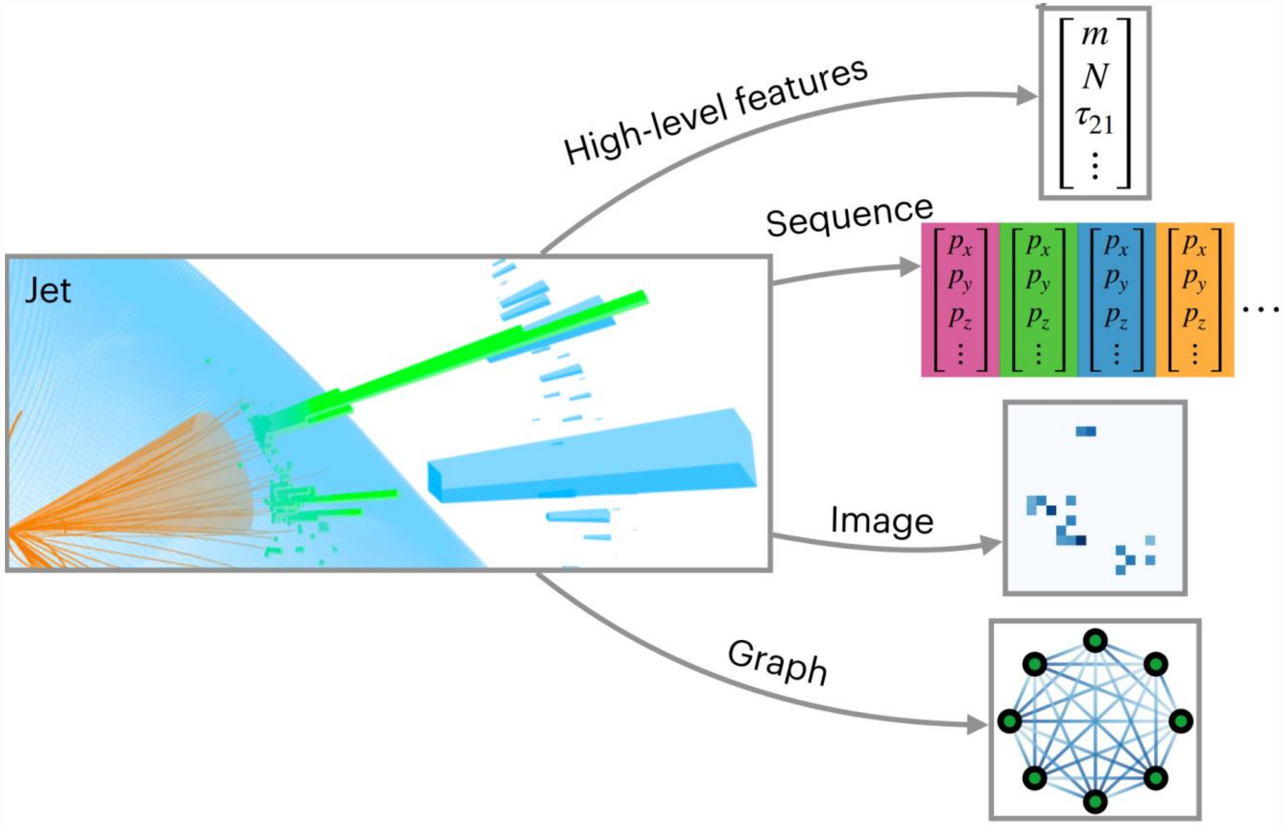
- ✓Classifier (Supervised)
- ✓Self-guided Detection/Search (Weakly Supervised/Unsupervised)
- ✓Reconstruction
- ✓Simulation
- ✓Language Model
- ✓Agent – an idea

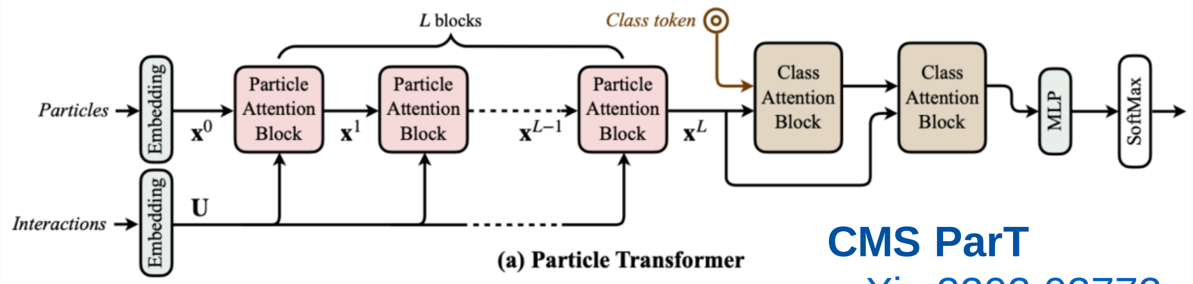
Smarter and More Sophisticated Classifier

Smarter and More Sophisticated Classifier



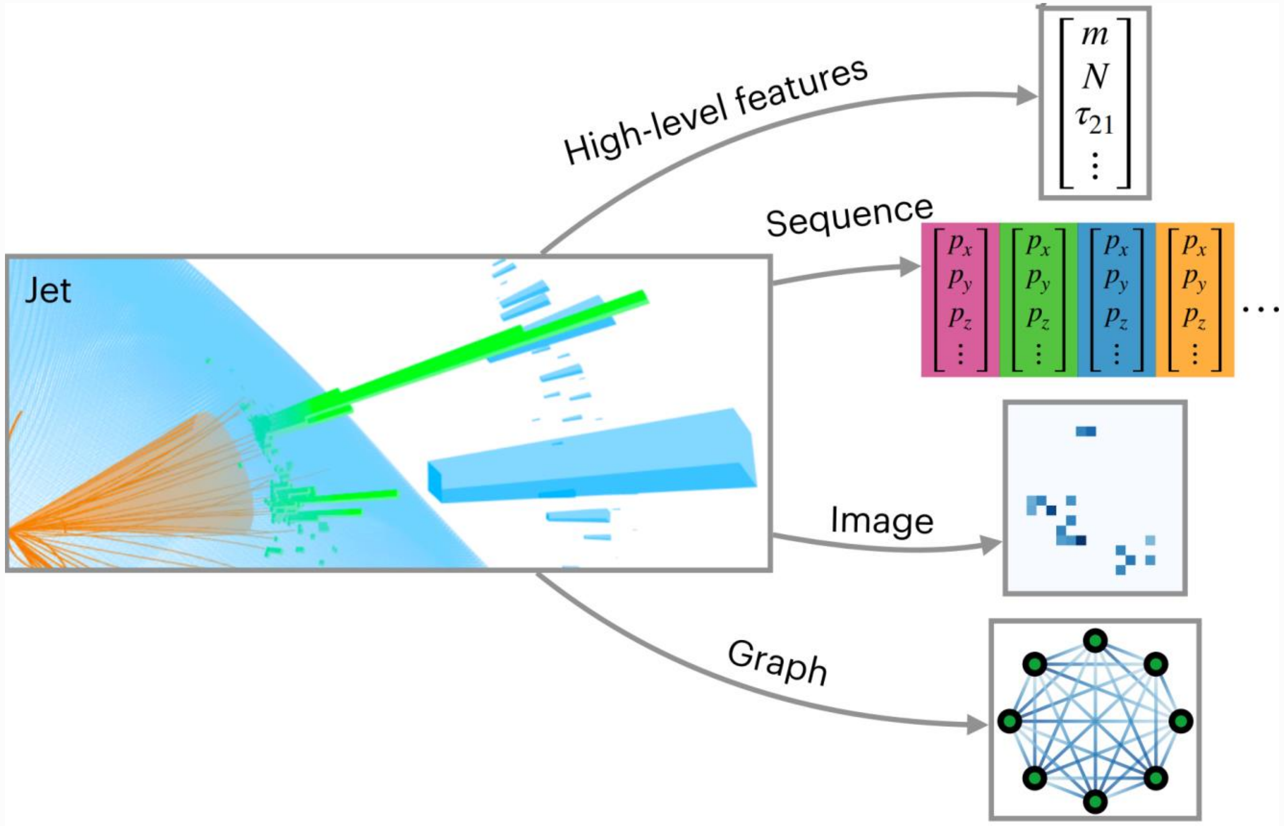
Smarter and More Sophisticated Classifier

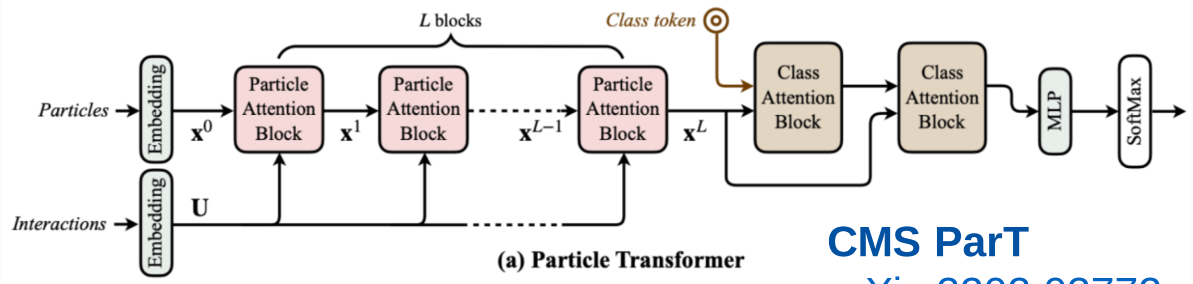




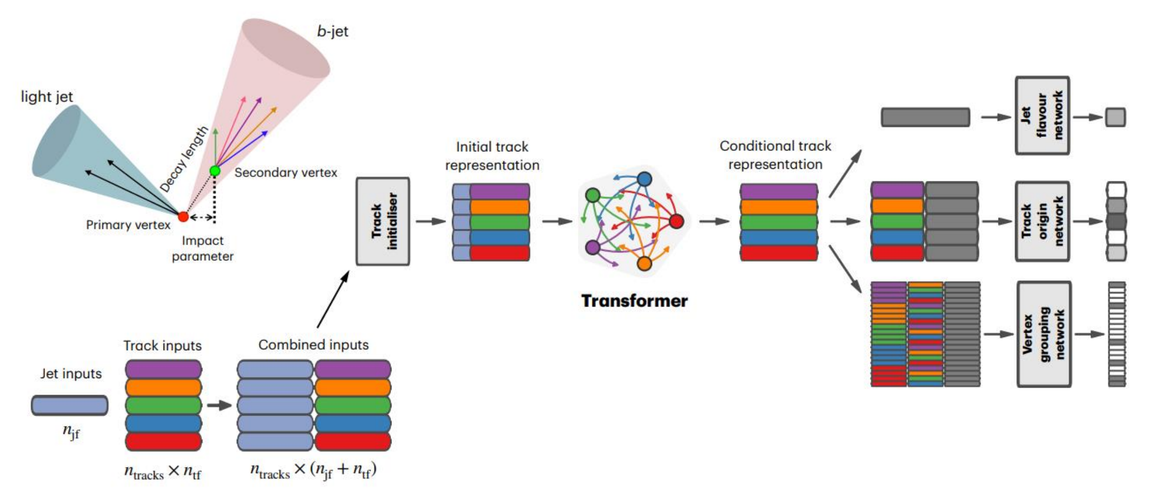
CMS ParT
arXiv:2202.03772

Smarter and More Sophisticated Classifier



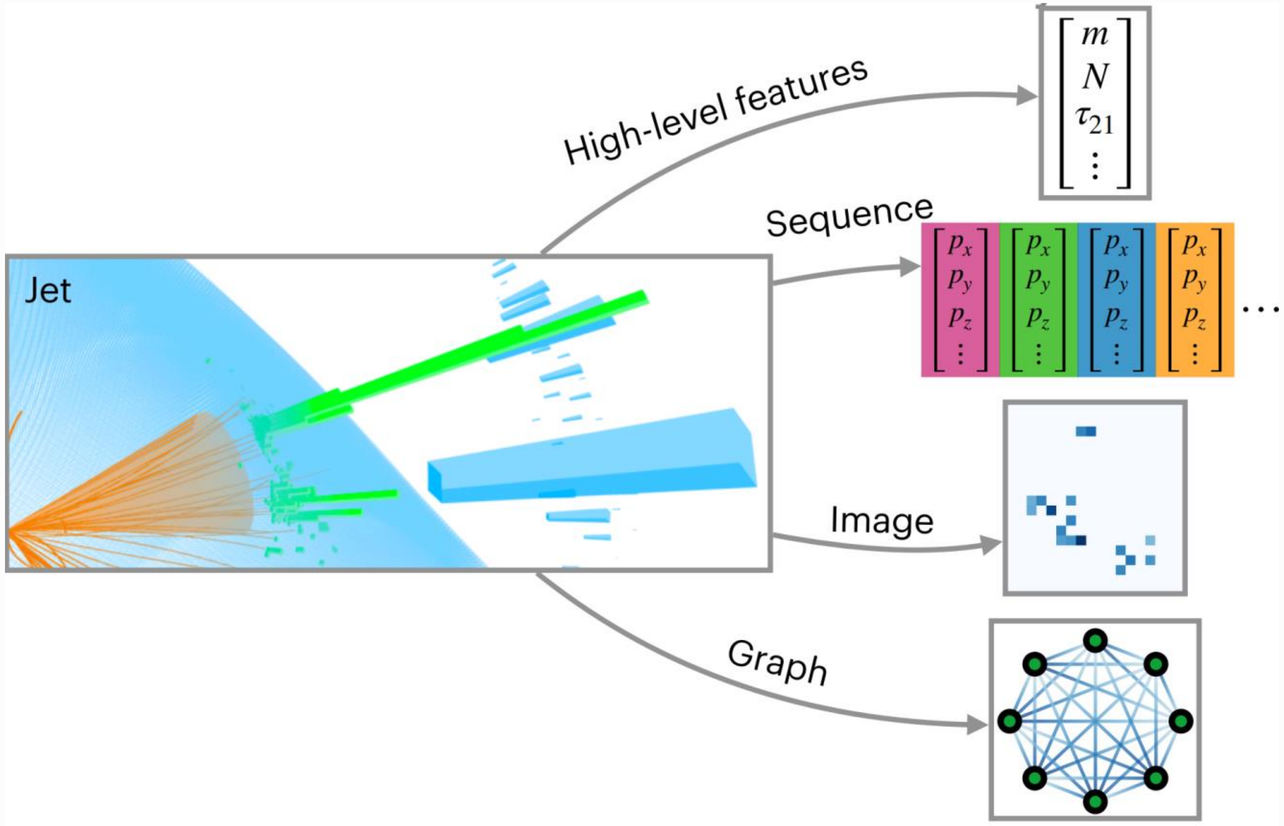


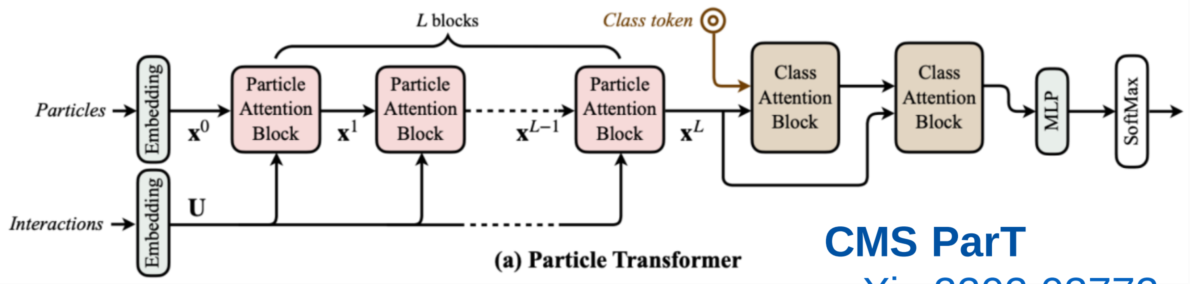
CMS ParT
arXiv:2202.03772



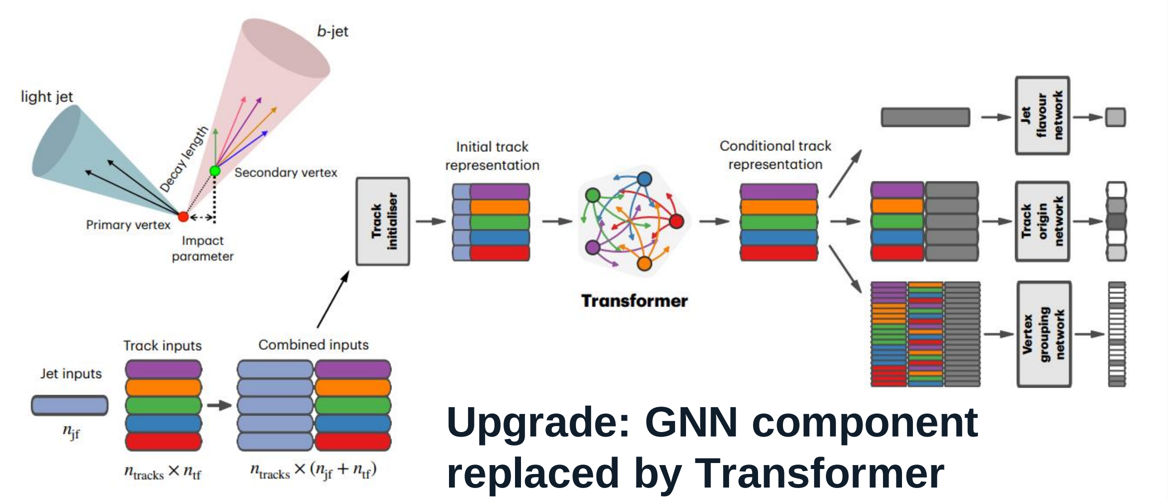
ATLAS General Network 2
arXiv:2505.19689

Smarter and More Sophisticated Classifier





CMS ParT
arXiv:2202.03772



Upgrade: GNN component replaced by Transformer

ATLAS General Network 2
arXiv:2505.19689

Smarter and More Sophisticated Classifier

Recap: Series of GloParT tagger (v1-v3)

Sep 2021

Initiated project: **ParticleNet for H→VV tagging** for novel boosted H→WW sea

JME presentations [22.01.11 by Cristina/Zichun][22.03.08 by Dawei][22.08.09 by Dawei

ParticleNet for HVV

Sep 2022

[22.12.13 by Congqiao (JME)]

- upgrade to **ParT architecture** (so-called GloParT v
- propose: "large model for large-scale classification" & subsequent fine-tuning capabilities (concept of Global ParT)

GloParT v1

[23.02.22 by Congqiao (ML Forum)]

the study of fine-tuning (f.t.) capabilities - **first demonstration**

May 2023

[23.07.24 by Congqiao (BTV)]

significantly improve pre-training comprehensiveness (extend to 314+2 nodes)

GloParT v2

[23.09.13 ML Town Hall] [23.12.05 ML Forum]

- re-studied f.t. capability with the stronger model
- study of f.t. for anomaly detection

July 2024

[24.09.20 by Congqiao (BTV)]

- v3 enhencement + extend to 374+374+2 nodes

GloParT v3

[24.10.18 Cross-POG]

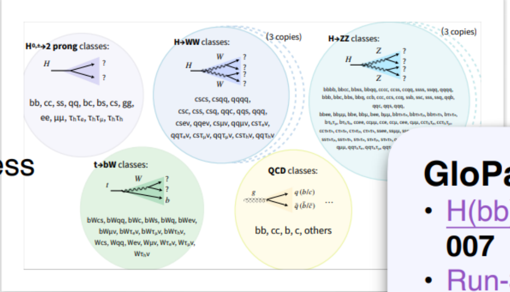
A comprehensive review prepared for GloParTv3's integration into cmssw

...

Process	Final state/ prongness	heavy flavour	# of classes
H→WW (full-hadronic)	qqqq	0c/1c/2c	3
	qqq		3
H→WW (semi-leptonic)	eνqq	0c/1c	2
	μνqq		2
	τ_eνqq		2
	τ_μνqq		2
	τ_hνqq		2
H→qq		bb	1
		cc	1
		ss	1
		qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$		1
	$\tau_\mu\tau_h$		1
	$\tau_h\tau_h$		1
t→bW (hadronic)	bqq	1b + 0c/1c	2
	bq		2
t→bW (leptonic)	beν	1b	1
	bμν		1
	bτ_eν		1
	bτ_μν		1
	bτ_hν		1
QCD		b	1
		bb	1
		c	1
		cc	1
		others (light)	1

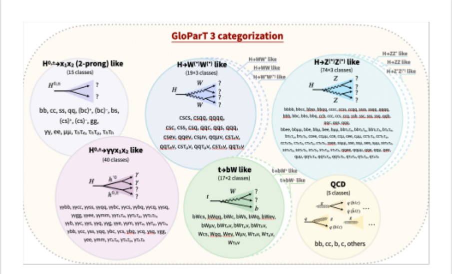
GloParT v1: used in the following analysis

- Boosted bbWW search: CMS-PAS-HIG-23-012
- HWW (0l/1l/VH): **HIG-24-008**
- X→H(bb)Y(WW): **B2G-23-007**



GloParT v2:

- H(bb)+γ: **B2G-24-007**
- Run-3 VH(bb/cc) (AK15): **HIG-25-001**
- boosted W→cb
- Run-3 HH(4b)



GloParT v3:

- Run-3 HH(4b)
- (more to join)

Global Particle Transformer (GloParT) algorithm

SJTUPA 上海交通大学物理与天文学院

Recap: Series of GloParT tagger (v1-v3)

Sep 2021

ParticleNet for HVV

Initiated project: **ParticleNet for H→VV tagging** for novel boosted H→WW sea

JME presentations[22.01.11 by Cristina/Zichun][22.03.08 by Dawei][22.08.09 by Dawe

Process	Final state/ prongness	heavy flavour	# of classes
H→WW (full-hadronic)	qqqq	0c/1c/2c	3
	qqq		3
H→WW (semi-leptonic)	eνqq	0c/1c	2
	μνqq		2
	$\tau_e\nu$qq		2
	$\tau_\mu\nu$qq		2
	$\tau_h\nu$qq		2
H→qq		bb	1
		cc	1
		ss	1
		qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$		1
	$\tau_\mu\tau_h$		1
	$\tau_h\tau_h$		1
t→bW (hadronic)	bqq	1b + 0c/1c	2
	bq		2
t→bW (leptonic)	beν	1b	1
	bμν		1
	b$\tau_e\nu$		1
	b$\tau_\mu\nu$		1
	b$\tau_h\nu$		1
QCD		b	1
		bb	1
		c	1
		cc	1
		others (light)	1

Sep 2022

GloParT v1

[22.12.13 by Congqiao (JME)]

- upgrade to **ParT architecture** (so-called GloParT v
- propose: "large model for large-scale classification" & subsequent fine-tuning capabilities (concept of Global ParT)

GloParT v1:
used in the following analysis
- Boosted bbWW search: **CMS-PAS-HIG-23-012**
- HWW (0l/1l/VH): **HIG-24-008**
- X→H(bb)Y(WW): **B2G-23-007**

[23.02.22 by Congqiao (ML Forum)]

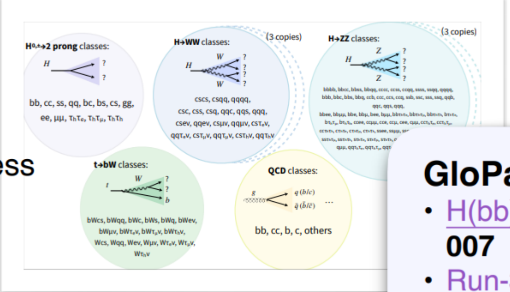
the study of fine-tuning (f.t.) capabilities - **first demonstration**

May 2023

GloParT v2

[23.07.24 by Congqiao (BTV)]

significantly improve pre-training comprehensiveness (extend to 314+2 nodes)



GloParT v2:
- H(bb)+γ: **B2G-24-007**
- Run-3 VH(bb/cc) (AK15): **HIG-25-001**
- boosted W→cb
- Run-3 HH(4b)

[23.09.13 ML Town Hall] [23.12.05 ML Forum]

- re-studied f.t. capability with the stronger model
- study of f.t. for anomaly detection

July 2024

GloParT v3

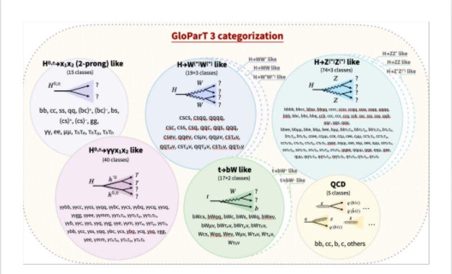
[24.09.20 by Congqiao (BTV)]

- v3 enhencement + extend to 374+374+2 nodes

[24.10.18 Cross-POG]

A comprehensive review prepared for GloParTv3's integration into cmssw

…



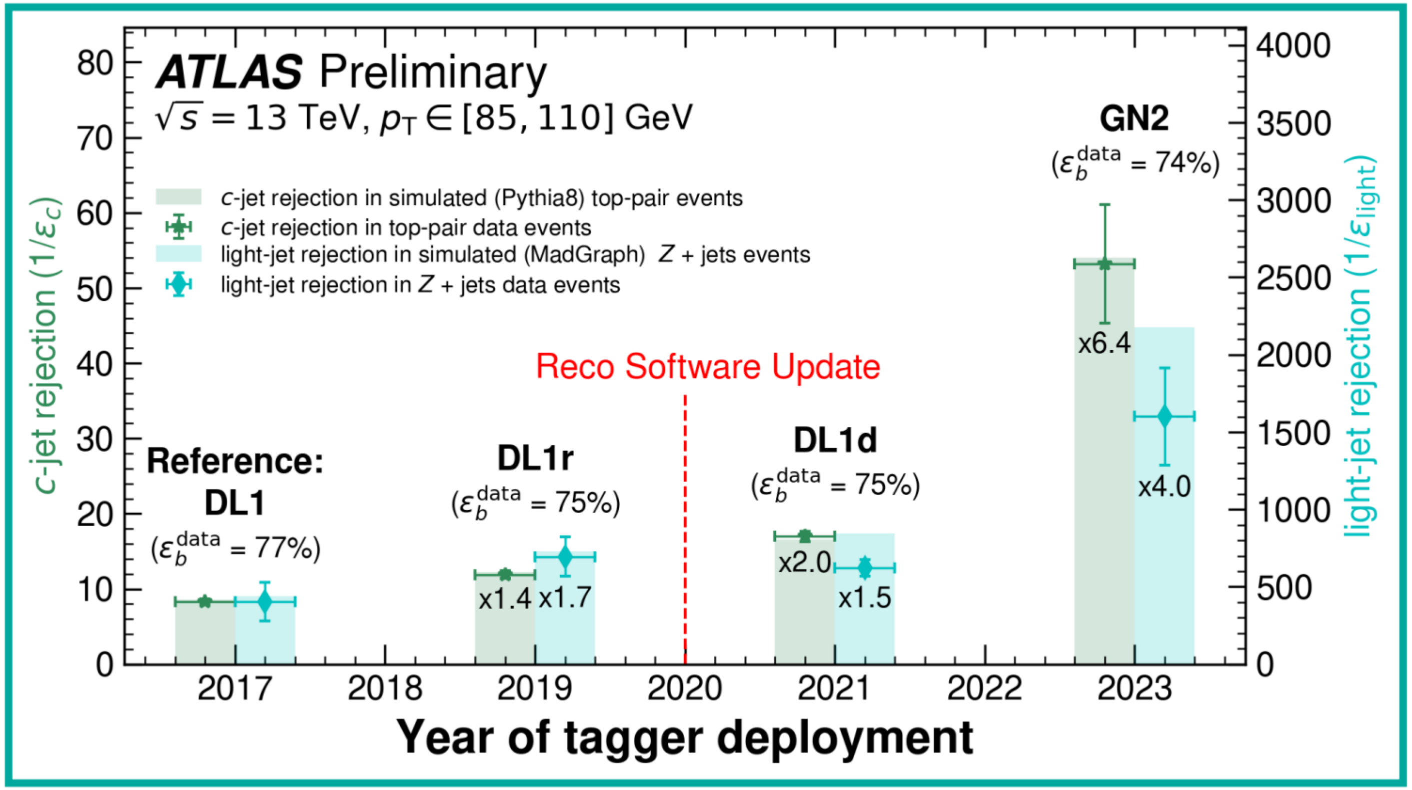
GloParT v3:
- Run-3 HH(4b)
- (more to join)

***Global Particle Transformer* (GloParT) algorithm**

Transformer model is based on "self-attention" mechanism: **Transfer model can focus on certain parts of the input data**, giving more weight to crucial features and disregarding unimportant ones.

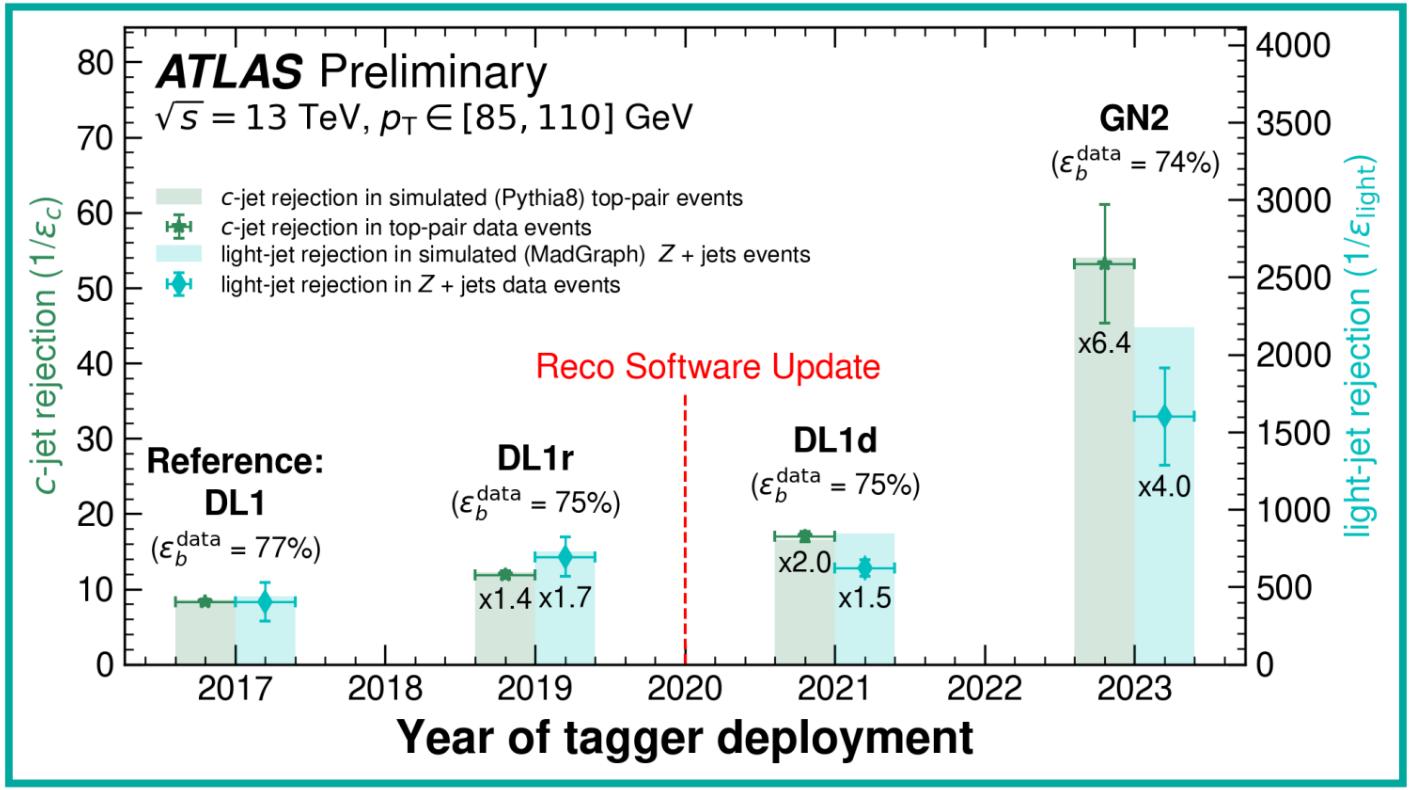
Better and More Robust Performance

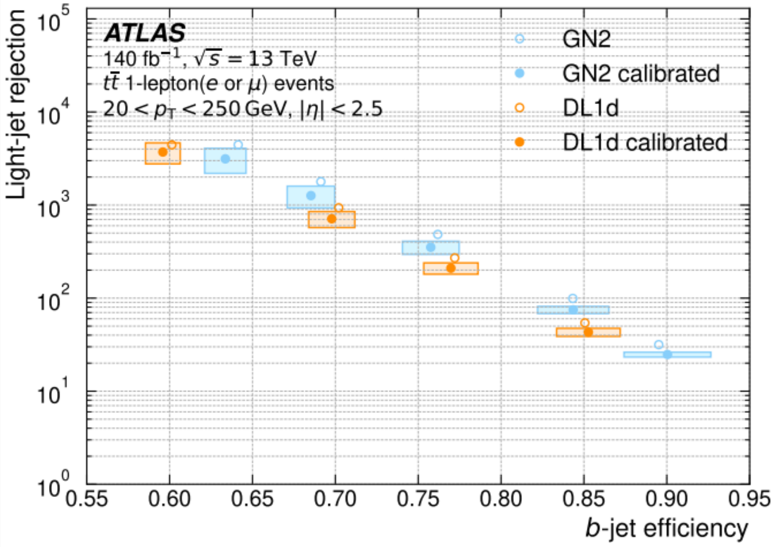
Better and More Robust Performance

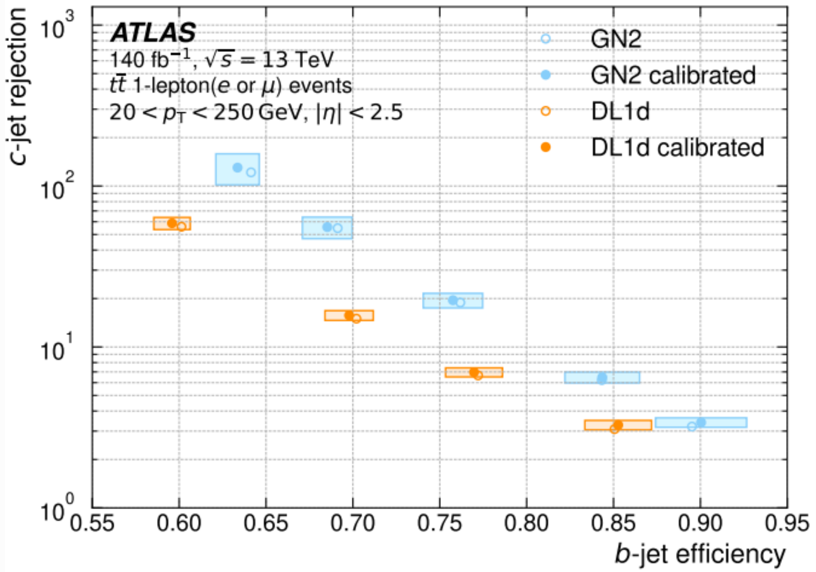


✓ **Significant improvement after years of development**

Better and More Robust Performance







- ✓ **Significant improvement after years of development**
- ✓ **Essential calibrations done for b-/c-jet and light jet flavors**
- ✓ **Performance in data matches simulation after calibration**

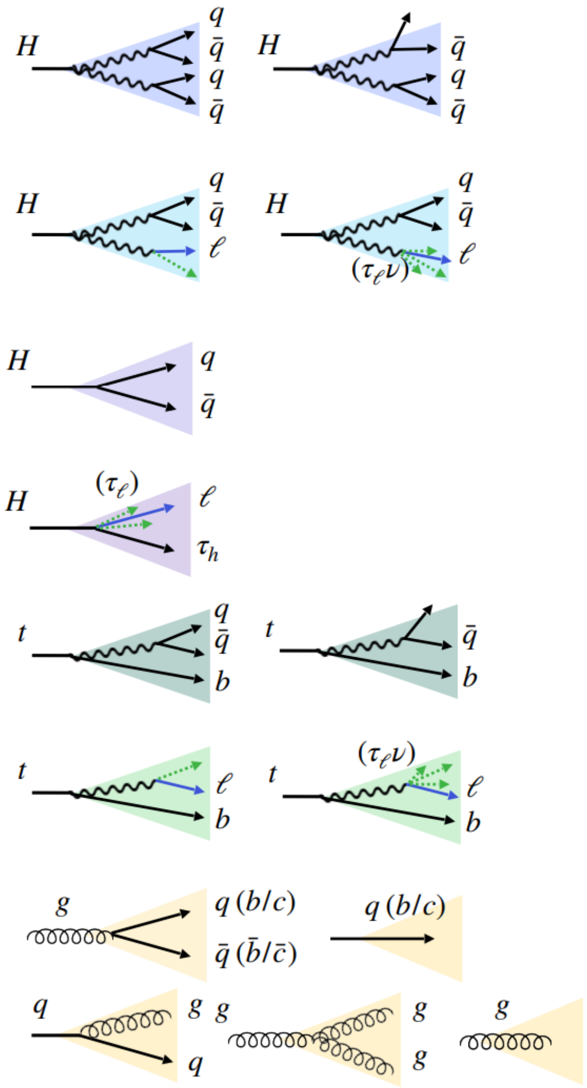
Better and More Robust Performance

CMS JME-25-001

Better and More Robust Performance

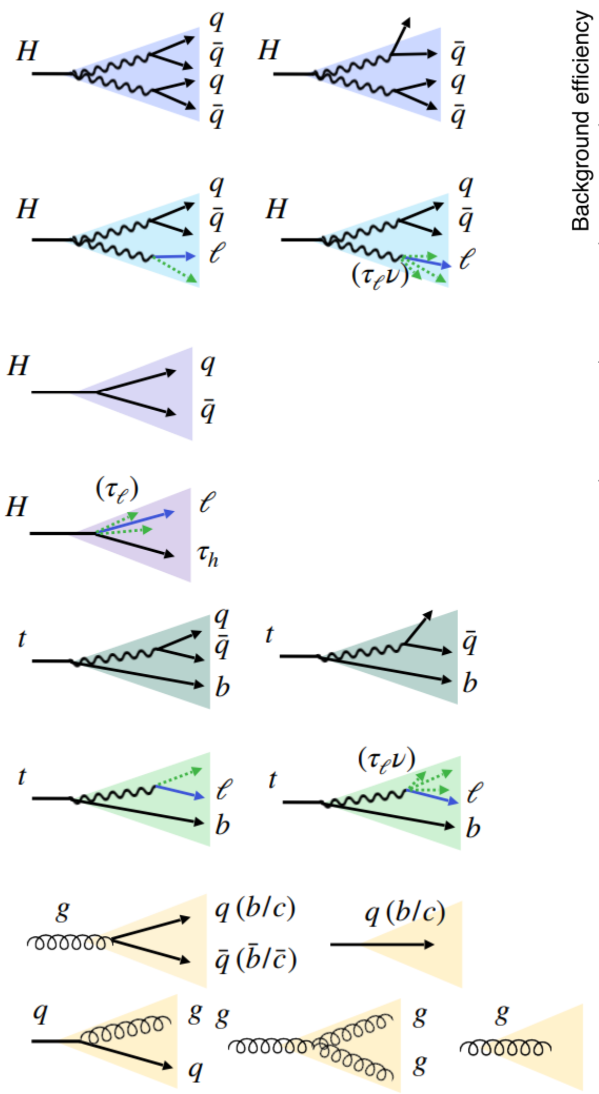
CMS JME-25-001

Process	Final state		Flavor	# of classes
H→WW (full-hadronic)	qqqq	⊗	0c / 1c / 2c	3
	qqq			3
H→WW (semi-leptonic)	eνqq	⊗	0c / 1c	2
	μνqq			2
	τ_eνqq			2
	τ_μνqq			2
	τ_hνqq			2
H→qq		⊗	bb	1
			cc	1
			ss	1
			qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$			1
	$\tau_\mu\tau_h$			1
	$\tau_h\tau_h$			1
t→bW (hadronic)	bqq	⊗	1b + 0c / 1c	2
	bq			2
t→bW (leptonic)	beν	⊗	1b	1
	bμν			1
	bτ_eν			1
	bτ_μν			1
	bτ_hν			1
QCD			b	1
			bb	1
			c	1
			cc	1
			others (light)	1

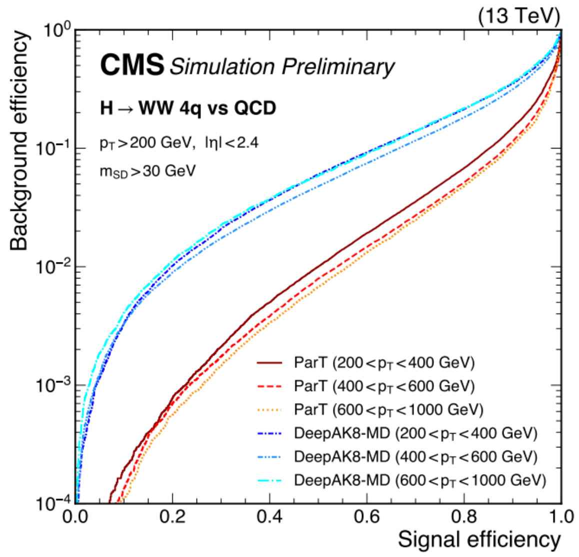


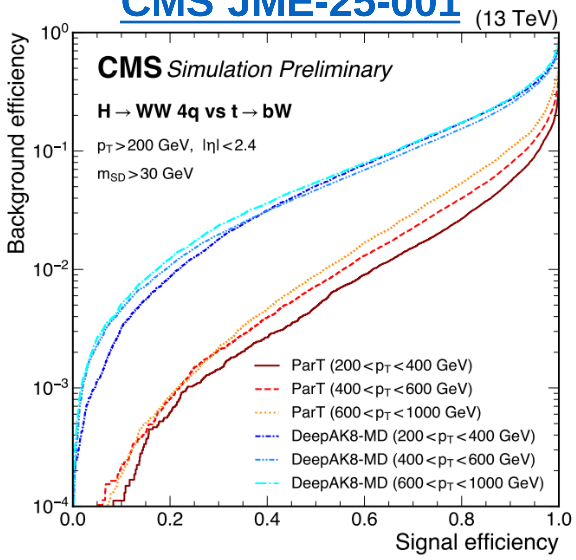
Better and More Robust Performance

Process	Final state		Flavor	# of classes
H→WW (full-hadronic)	qqqq	⊗	0c / 1c / 2c	3
	qqq			3
H→WW (semi-leptonic)	eνqq	⊗	0c / 1c	2
	μνqq			2
	τ_eνqq			2
	τ_μνqq			2
	τ_hνqq			2
H→qq		⊗	bb	1
			cc	1
			ss	1
			qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$			1
	$\tau_\mu\tau_h$			1
	$\tau_h\tau_h$			1
t→bW (hadronic)	bqq	⊗	1b + 0c / 1c	2
	bq			2
t→bW (leptonic)	beν	⊗	1b	1
	bμν			1
	bτ_eν			1
	bτ_μν			1
	bτ_hν			1
QCD			b	1
			bb	1
			c	1
			cc	1
			others (light)	1



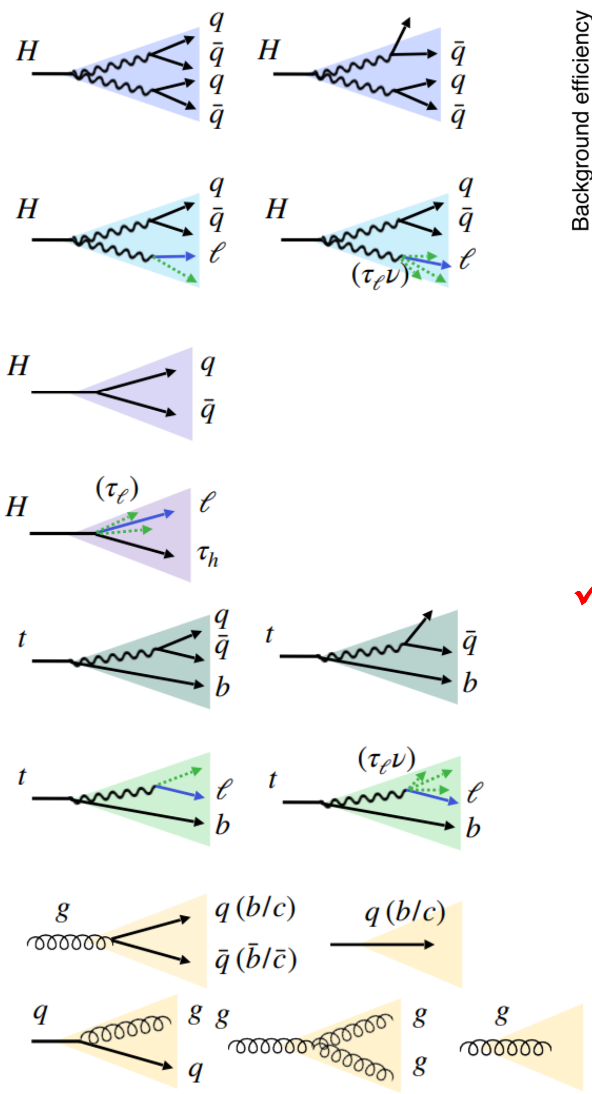
CMS JME-25-001



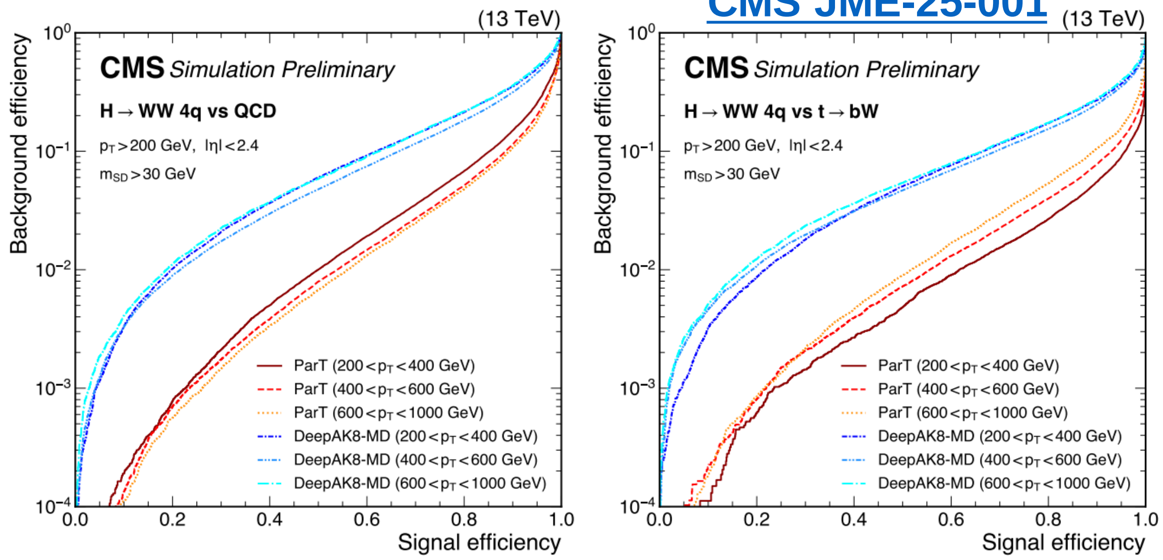


Better and More Robust Performance

Process	Final state		Flavor	# of classes
H→WW (full-hadronic)	qqqq	⊗	0c / 1c / 2c	3
	qqq			3
H→WW (semi-leptonic)	eνqq	⊗	0c / 1c	2
	μνqq			2
	τ_eνqq			2
	τ_μνqq			2
	τ_hνqq			2
H→qq		⊗	bb	1
			cc	1
			ss	1
			qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$			1
	$\tau_\mu\tau_h$			1
	$\tau_h\tau_h$			1
t→bW (hadronic)	bqq	⊗	1b + 0c / 1c	2
	bq			2
t→bW (leptonic)	beν	⊗	1b	1
	bμν			1
	bτ_eν			1
	bτ_μν			1
	bτ_hν			1
QCD			b	1
			bb	1
			c	1
			cc	1
			others (light)	1



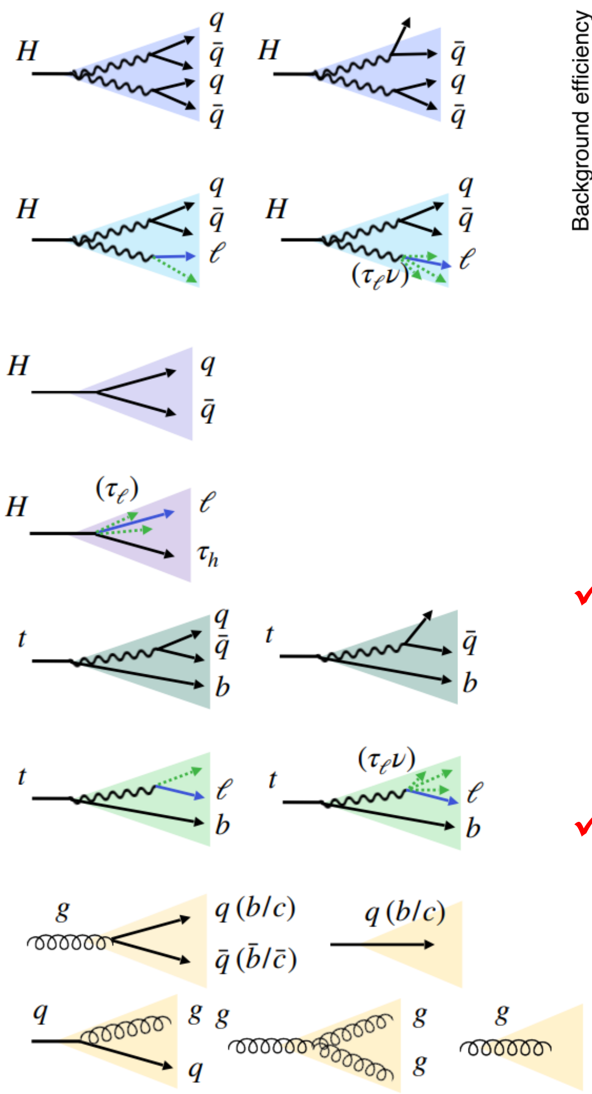
CMS JME-25-001



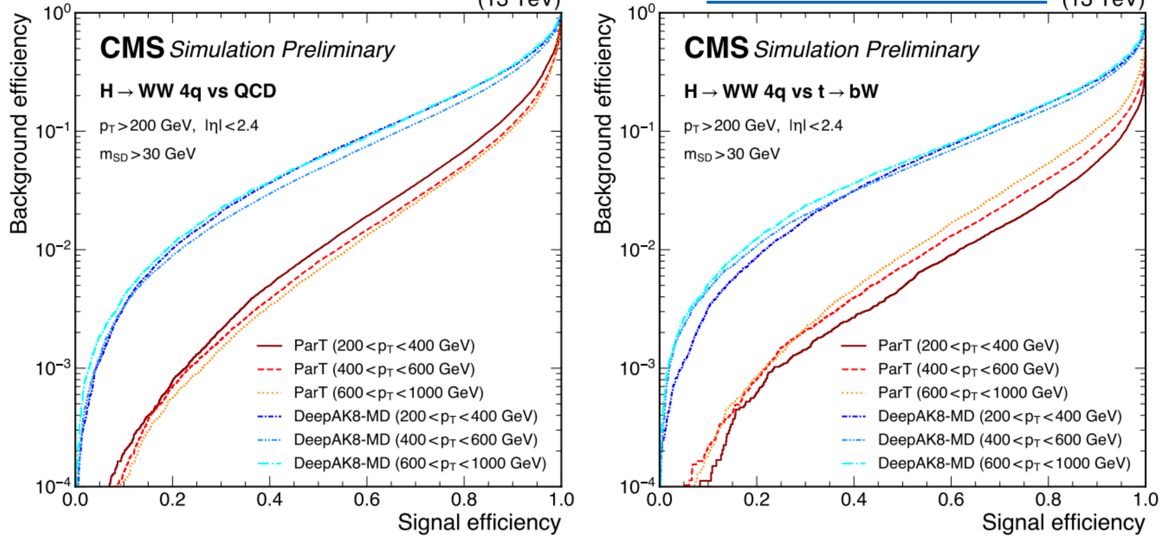
- ✓ **Highly granular multi-classifier gives 6-20 fold improvement in background rejection rate on H→WW* →4j vs. QCD/top jets**
  - **Compared with early DeepAK8-MD tagger**

Better and More Robust Performance

Process	Final state		Flavor	# of classes
H→WW (full-hadronic)	qqqq	⊗	0c / 1c / 2c	3
	qqq			3
H→WW (semi-leptonic)	eνqq	⊗	0c / 1c	2
	μνqq			2
	τ_eνqq			2
	τ_μνqq			2
	τ_hνqq			2
H→qq		⊗	bb	1
			cc	1
			ss	1
			qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$			1
	$\tau_\mu\tau_h$			1
	$\tau_h\tau_h$			1
t→bW (hadronic)	bqq	⊗	1b + 0c / 1c	2
	bq			2
t→bW (leptonic)	beν	⊗	1b	1
	bμν			1
	bτ_eν			1
	bτ_μν			1
	bτ_hν			1
QCD			b	1
			bb	1
			c	1
			cc	1
			others (light)	1



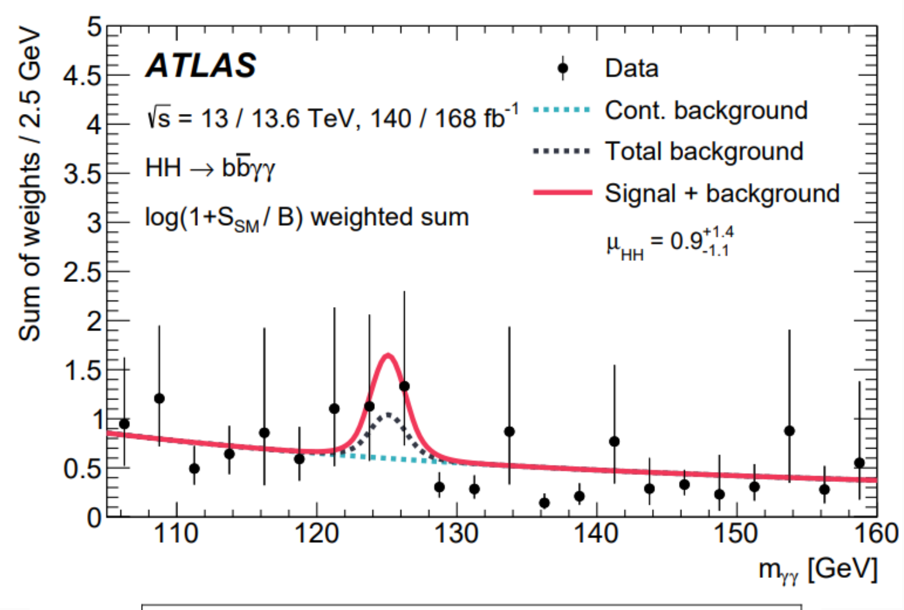
CMS JME-25-001

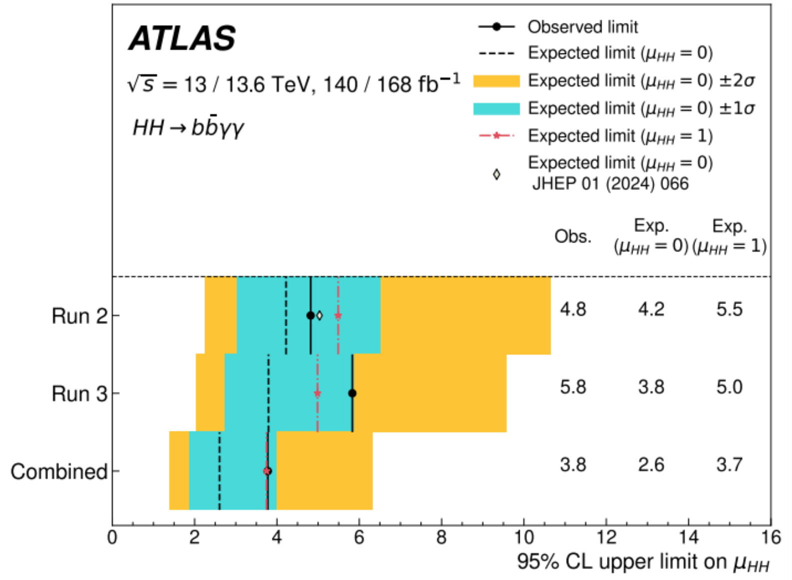


- ✓ **Highly granular multi-classifier gives 6-20 fold improvement in background rejection rate on H→WW* →4j vs. QCD/top jets**
  - **Compared with early DeepAK8-MD tagger**
- ✓ **Challenge for tagger calibrations**
  - **Hard to find SM events in similar topology**
  - **New technique uses Lund jet plane**
  - **Effectively measure scale factors per quark sub-jet**

Impact on Physics

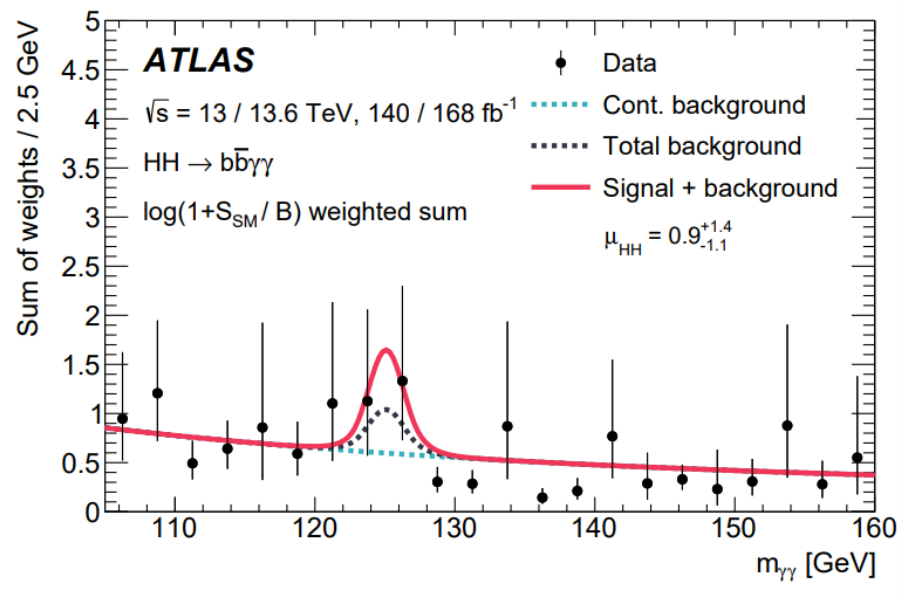
Impact on Physics

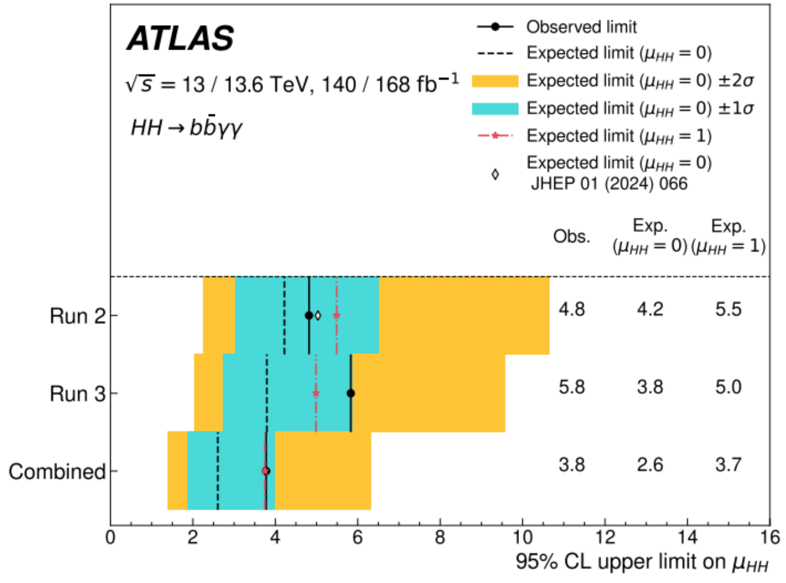




GN2 alone brings 20% improvement for $HH \rightarrow bb\gamma\gamma$ analysis

arXiv:2507.03495

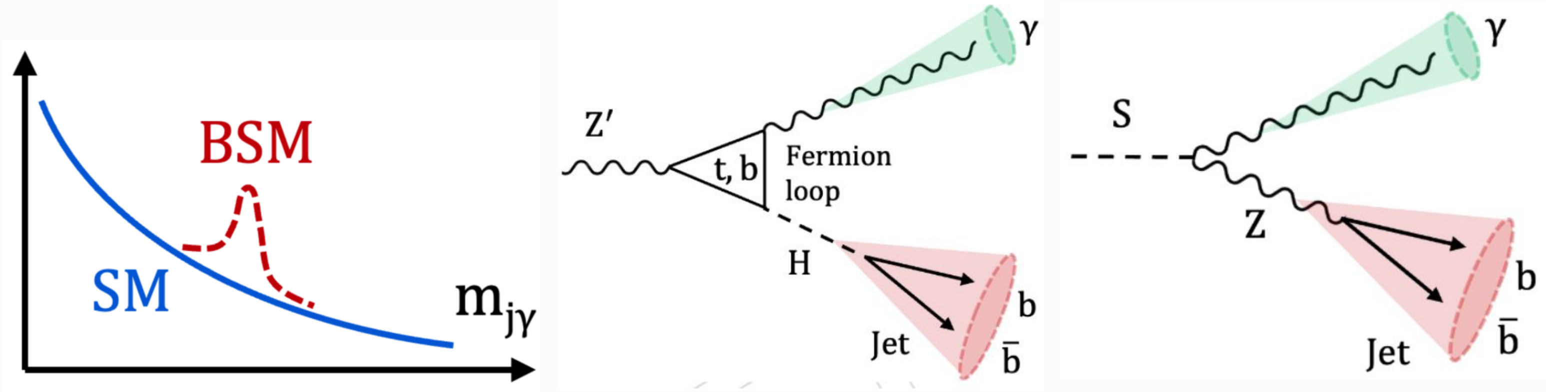


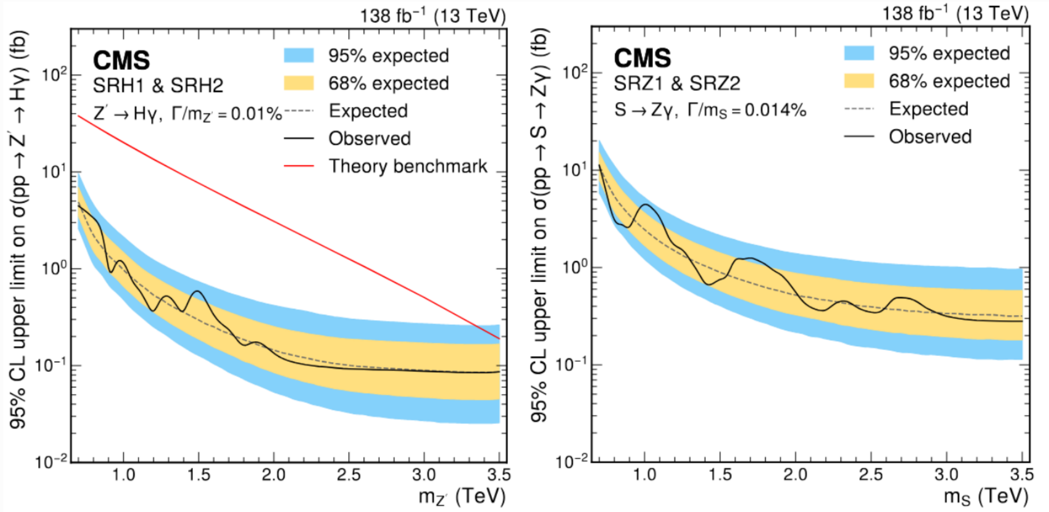


GN2 alone brings 20% improvement for HH → bbγγ analysis

[arXiv:2507.03495](https://arxiv.org/abs/2507.03495)

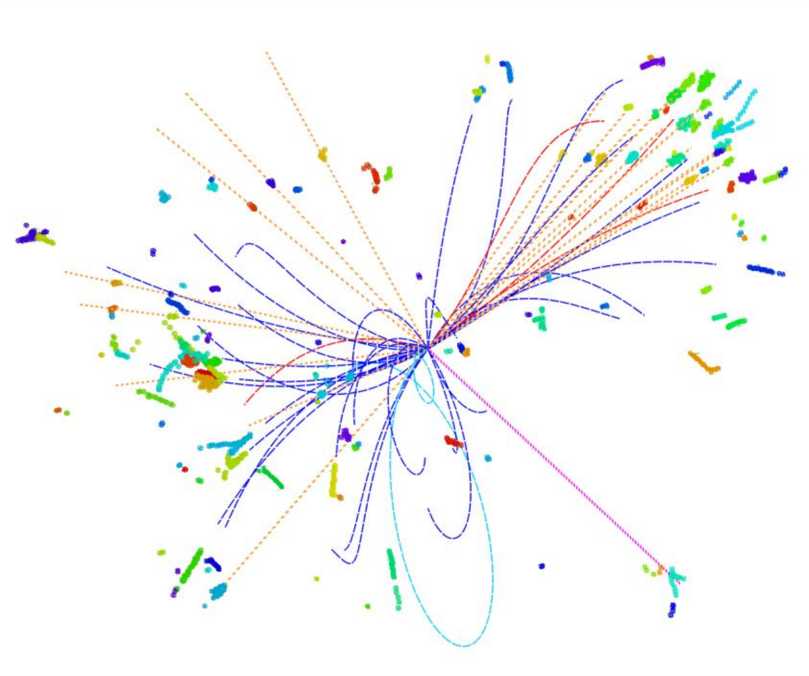
High-mass resonances in H/Z(bb)+γ final state



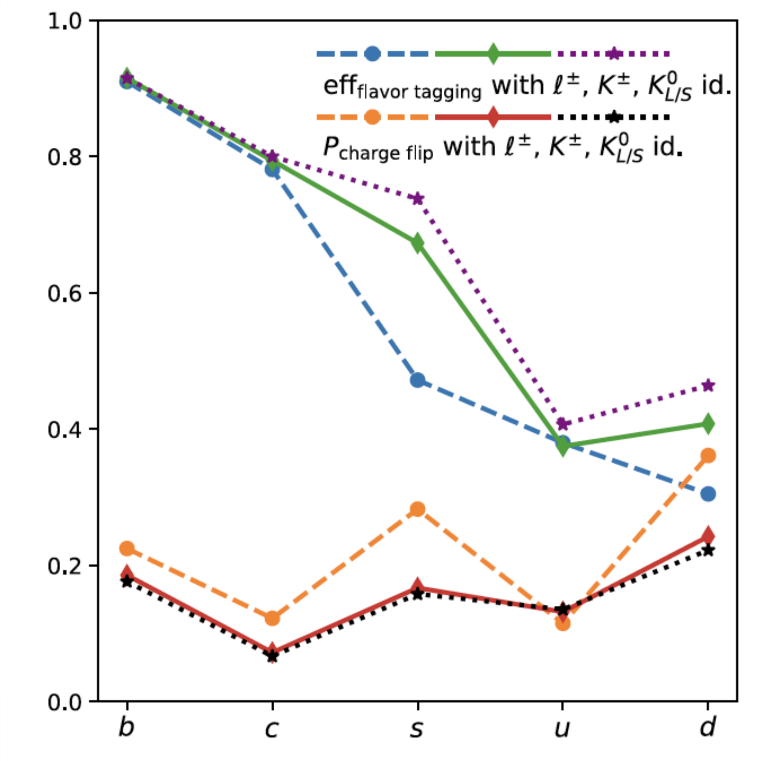


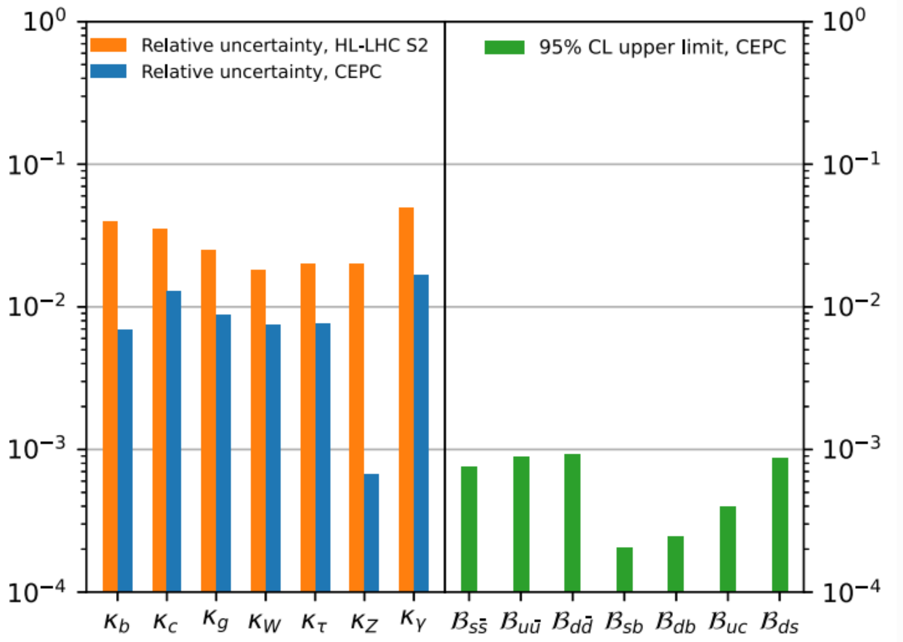
- ✓ **GloParT V2 used for X→bb tagger**
  - **H/Z→bb vs. QCD jets**
- ✓ **Most stringent limits for both channels**

Impact on Physics



$e^+e^- \rightarrow \nu\bar{\nu}H \rightarrow \nu\bar{\nu}gg$





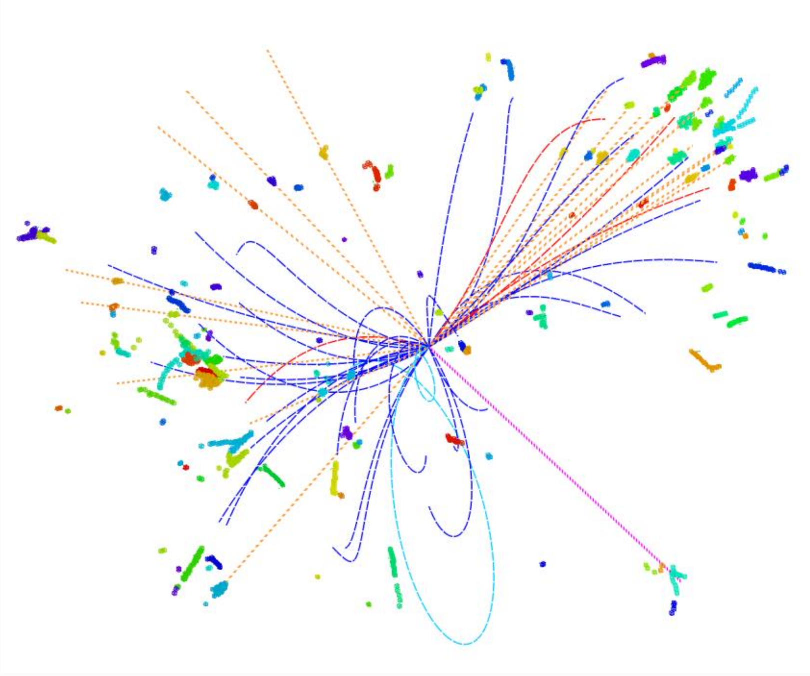
PRL 132, 221802 (2024)

Jet Origin ID:

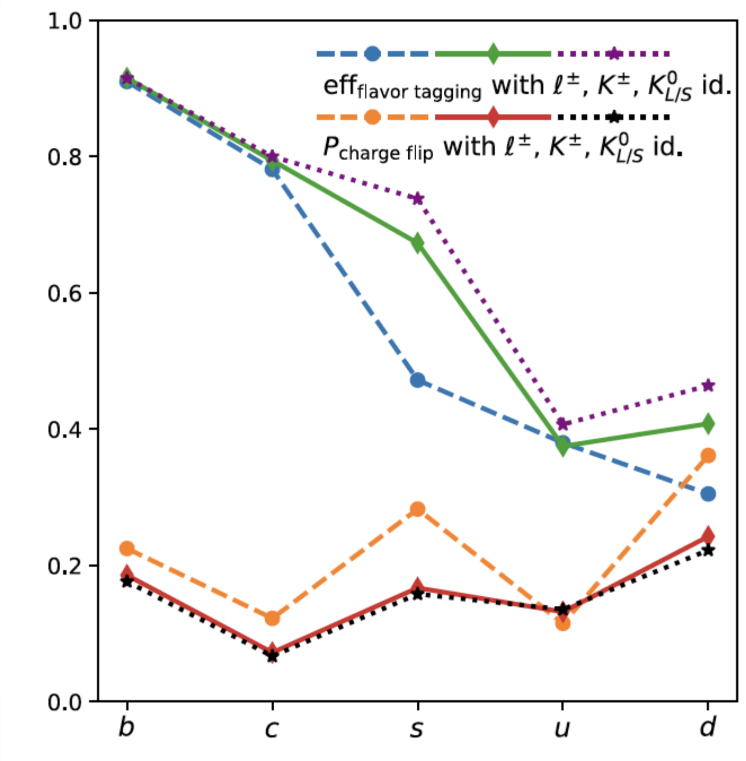
- **11 categories (5 quarks + 5 anti quarks + gluon) identification, realized at Full Simulated di-jet events at CEPC CDR baseline with Arbor + ParticleNet (GNN).**
- **Jet flavor tagging efficiencies ranging from 67% to 92% for b-, c-, and s-quarks and jet charge flip rates of 7%–24% for all quark species. Higgs decay BRs range from 2×10^{-4} to 1×10^{-3} (95% C.L.).**

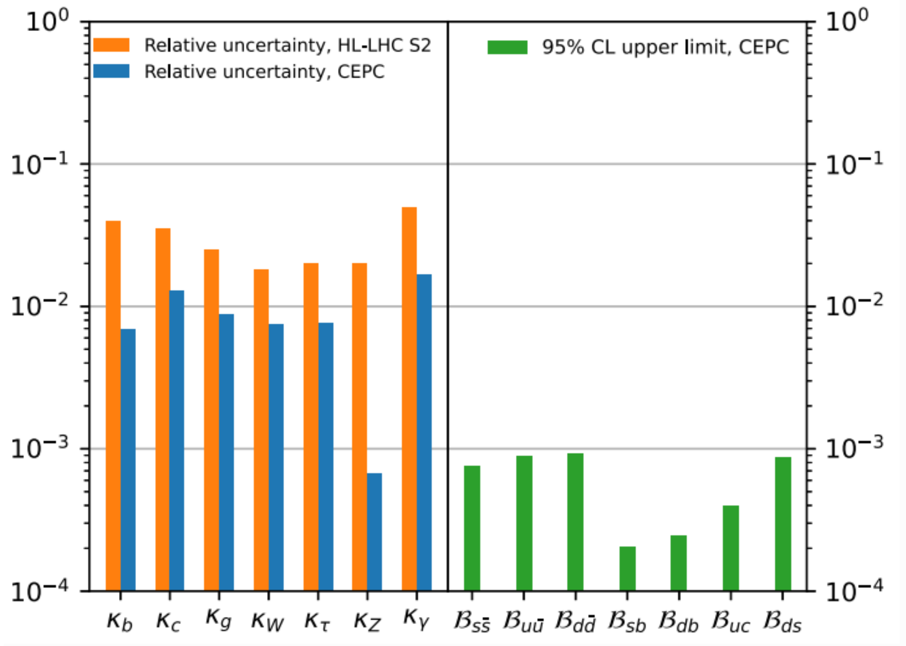
Impact on Physics

✓ See more in Manqi's talk on Friday



$e^+e^- \rightarrow \nu\bar{\nu}H \rightarrow \nu\bar{\nu}gg$





PRL 132, 221802 (2024)

Jet Origin ID:

- **11 categories (5 quarks + 5 anti quarks + gluon) identification, realized at Full Simulated di-jet events at CEPC CDR baseline with Arbor + ParticleNet (GNN).**
- **Jet flavor tagging efficiencies ranging from 67% to 92% for b-, c-, and s-quarks and jet charge flip rates of 7%–24% for all quark species. Higgs decay BRs range from 2×10^{-4} to 1×10^{-3} (95% C.L.).**

Let’s Go Even Bigger: Event-Level Foundation Model

Core Idea: One strong body + many small heads

🧠 Decoder – Discriminative Heads：

Segmentation

- Inspired by Meta AI’s segmentation networks
  - The model performs set prediction (queries → predict class & mask), preserving permutation symmetry.
  - Naturally extendable from objects to substituents without changing the model design.

Classification

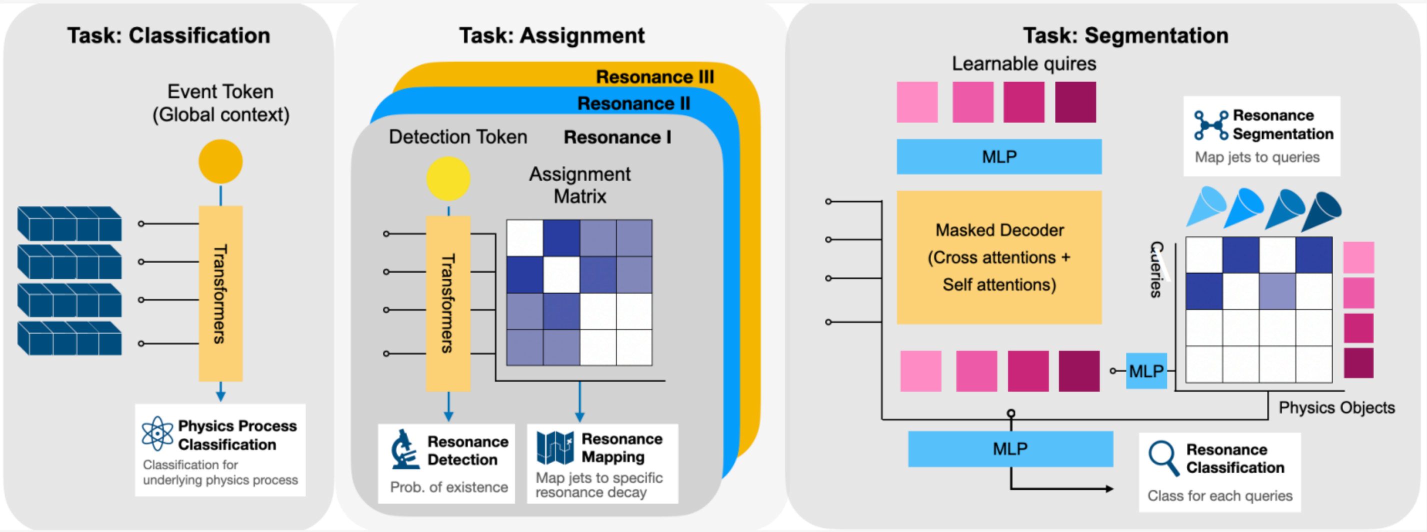
- Multi-Class event classifiers (with regression)

Assignment

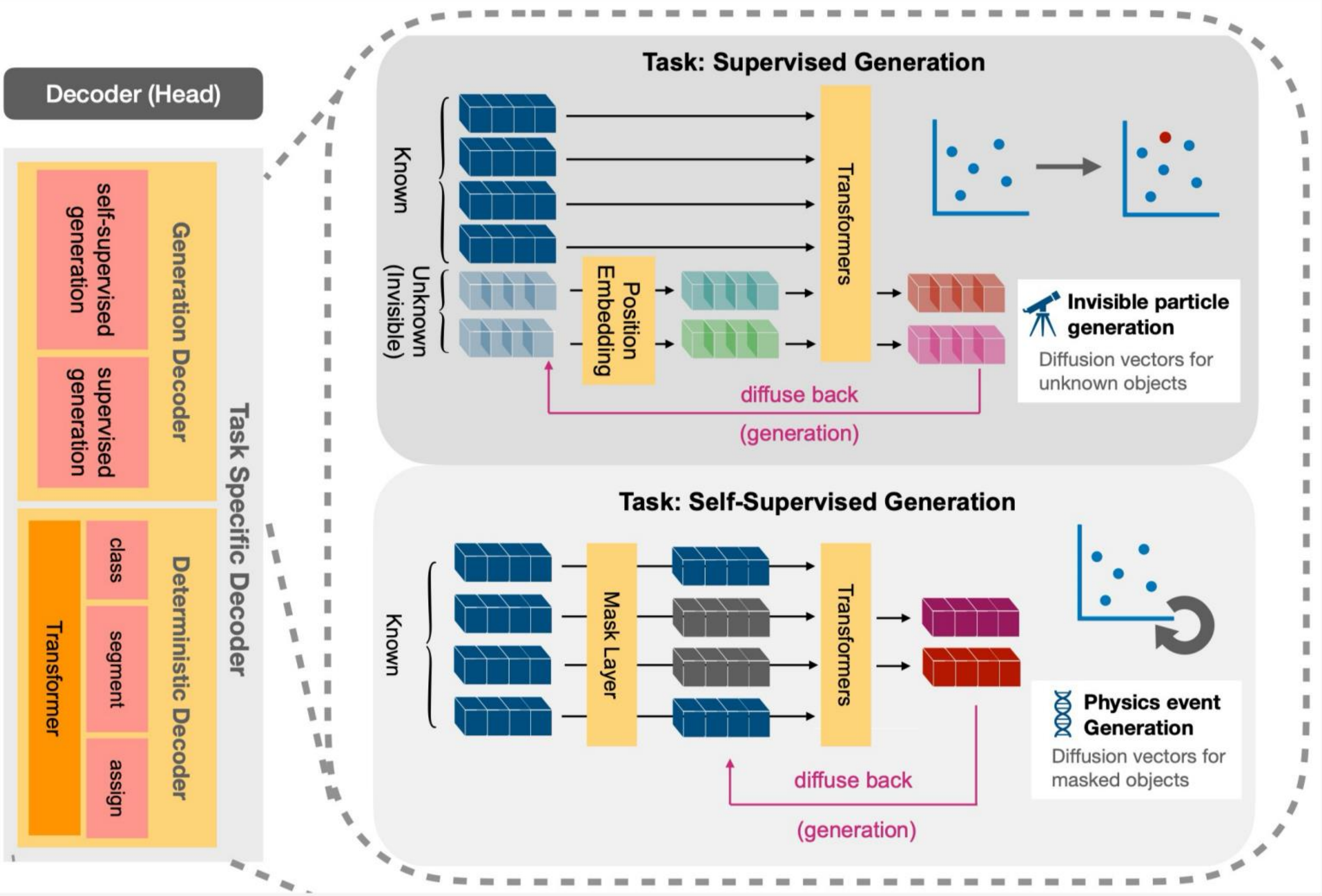
- Symmetry-aware mapping of objects to truth partons (requires known decay topology).
- High accuracy for well-defined processes, but rigid, costly, not generalizable.

Input Representation

- 📌 **Particle Cloud (Up to 18 Particles per Event):**
  - Each particle is encoded with 7 features: **4-momentum, isbJet, isLepton**, and **charge.**
- 🌐 **Global Features / Event Observables:**
  - Missing transverse energy
  - Number of leptons, number of jets
  - Invariant mass of visible objects
  - Scalar sums like **HT, ST**, etc.



EveNet: Event-Level Foundation Model



Core Idea: One strong body + many small heads

🧠 Decoder – Generation Head：

Supervised Generation

- Use known objects as input to predict missing ones (e.g., neutrinos).
- Diffusion models capture high-dimensional probability densities → predict the most likely kinematics.

Self-supervised Generation

- Mask part/all of the inputs and reconstruct them with a diffusion model.
- Learns underlying event structure without requiring labels.

Weakly Supervised: Search for Exotic Higgs Decay

✓ **Jointly trained on the Assignment and Classification tasks**

2-15x improvement on bkgd. rejection

- **Signal:** $H \rightarrow aa \rightarrow bbbb$

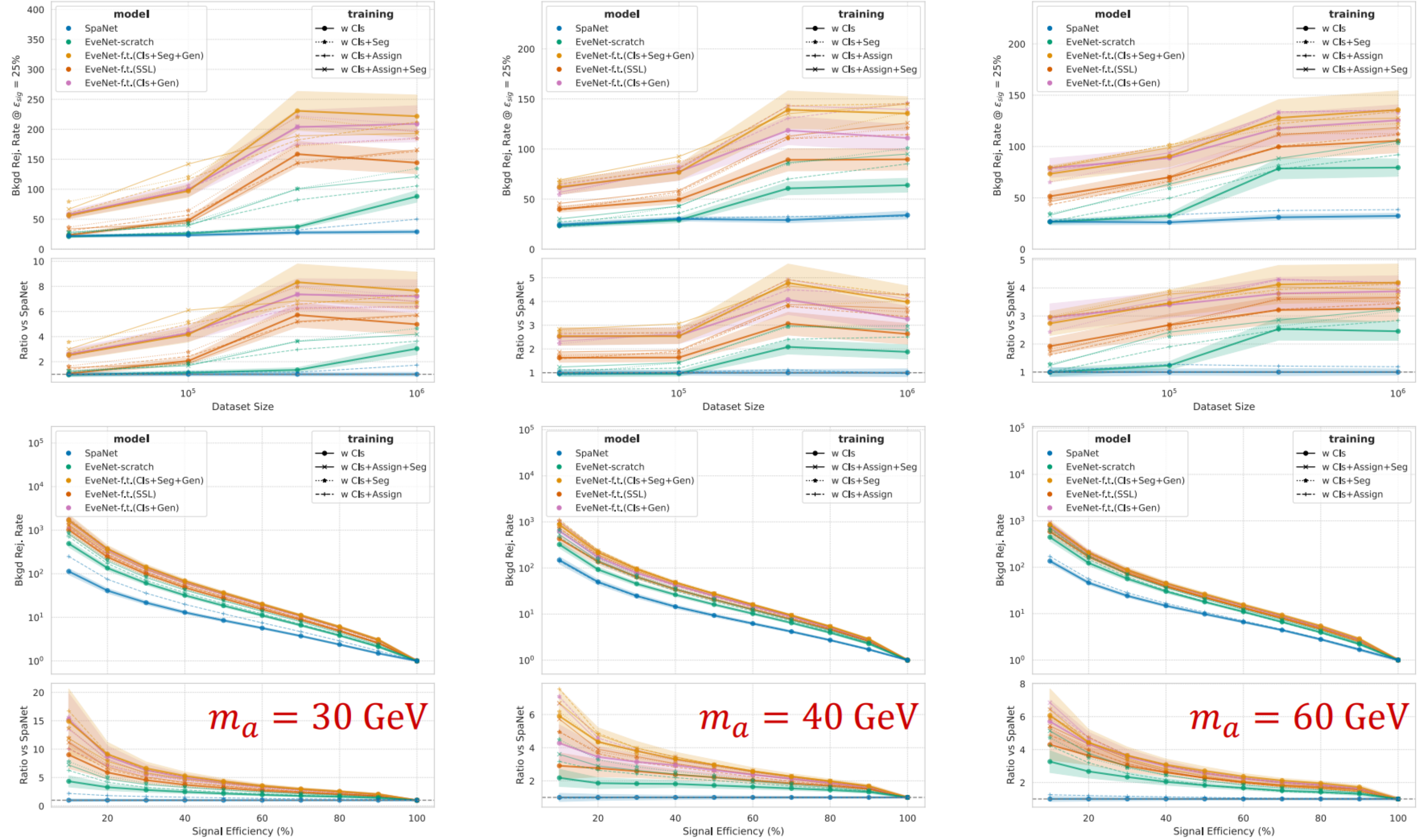
$(m_a = 30, 40, 60\ \text{GeV})$

- **QCD:** $bbbb, bbbj, bbjj$

- **Reference Network:** SPANet (same hidden dim)

Classification

1. **Inversed ROC**
2. **Bkgd. Rejection rate @ signal efficiency of 25%**



Weakly Supervised: Search for Exotic Higgs Decay

SJTUPA 上海交通大学物理与天文学院

✓ **Jointly trained on the Assignment and Classification tasks**

2-15x improvement on bkgd. rejection

- **Signal:** $H \rightarrow aa \rightarrow bbbb$

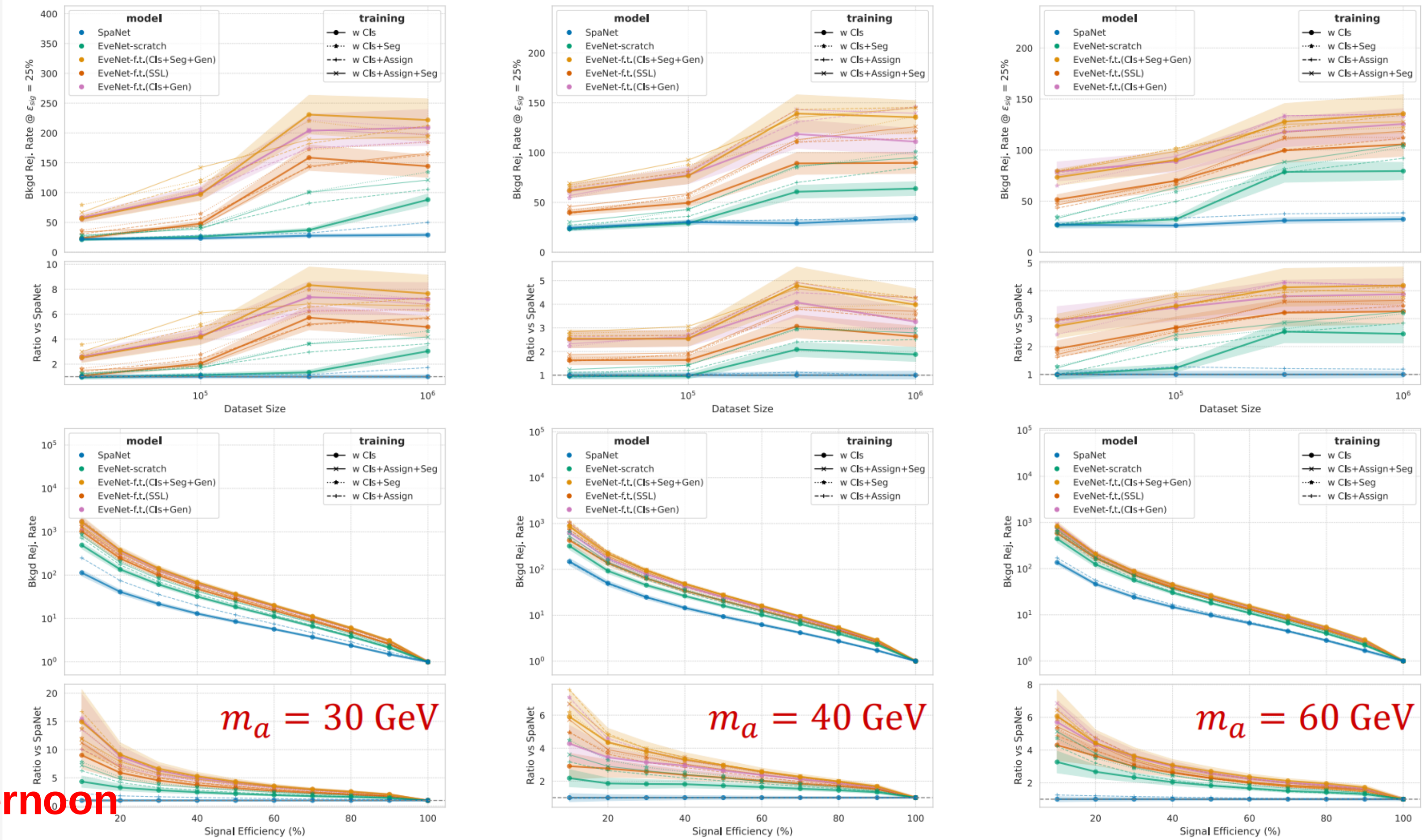
$(m_a = 30, 40, 60 \text{ GeV})$

- **QCD:** $bbbb, bbbj, bbjj$
- **Reference Network:** SPANet

(same hidden dim)

Classification

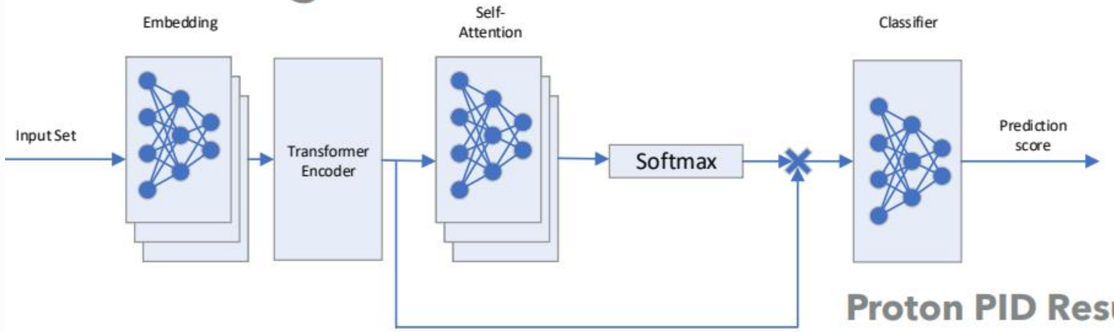
1. **Inversed ROC**
2. **Bkgd. Rejection rate @ signal efficiency of 25%**

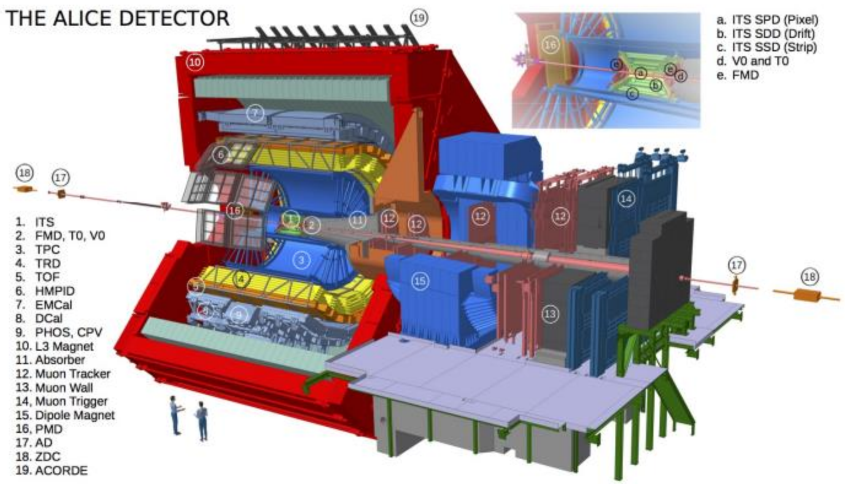


✓ **See more in Yulei's talk in the afternoon**

Reconstruction: PID

- **Transformer** for particle ID in ALICE can result in higher purity and efficiency than standard methods
- Use **domain adversarial neural networks** to mitigate data-simulation differences

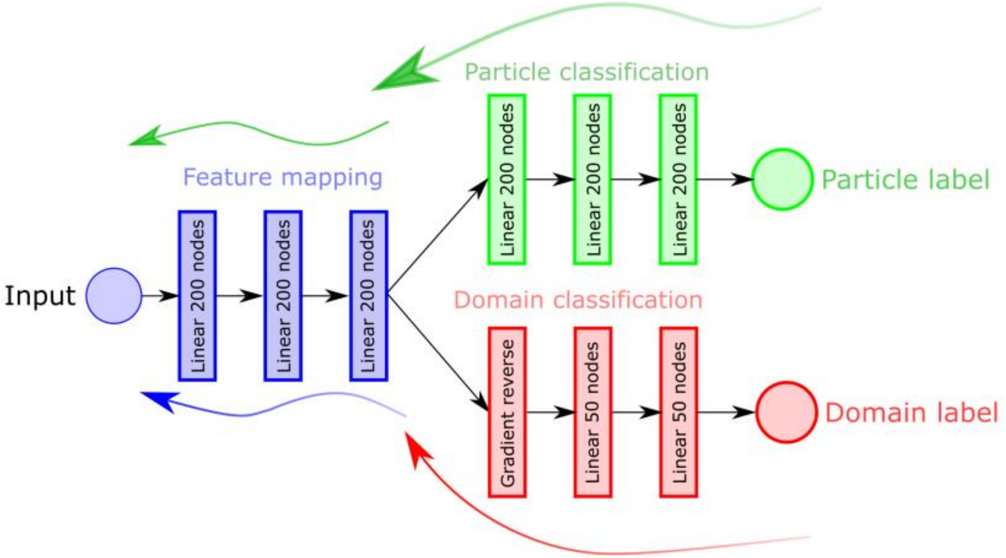




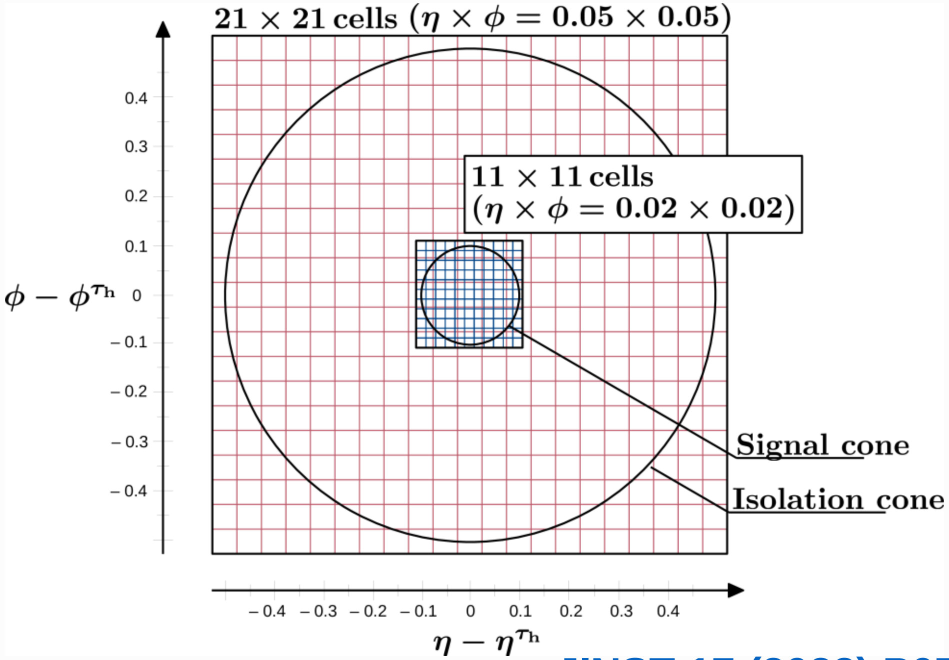
Proton PID Results

Model	Precision	Recall	F_1
Standard	**99.40 ± 0.01**	59.72 ± 0.03	74.61 ± 1.88
Ensemble	97.16 ± 0.46	93.74 ± 0.30	95.42 ± 0.12
Mean	97.85 ± 0.41	93.34 ± 0.32	95.54 ± 0.06
Proposed	97.80 ± 0.44	**93.86 ± 0.27**	**95.79 ± 0.07**
Regression	97.38 ± 0.40	93.67 ± 0.38	95.49 ± 0.15

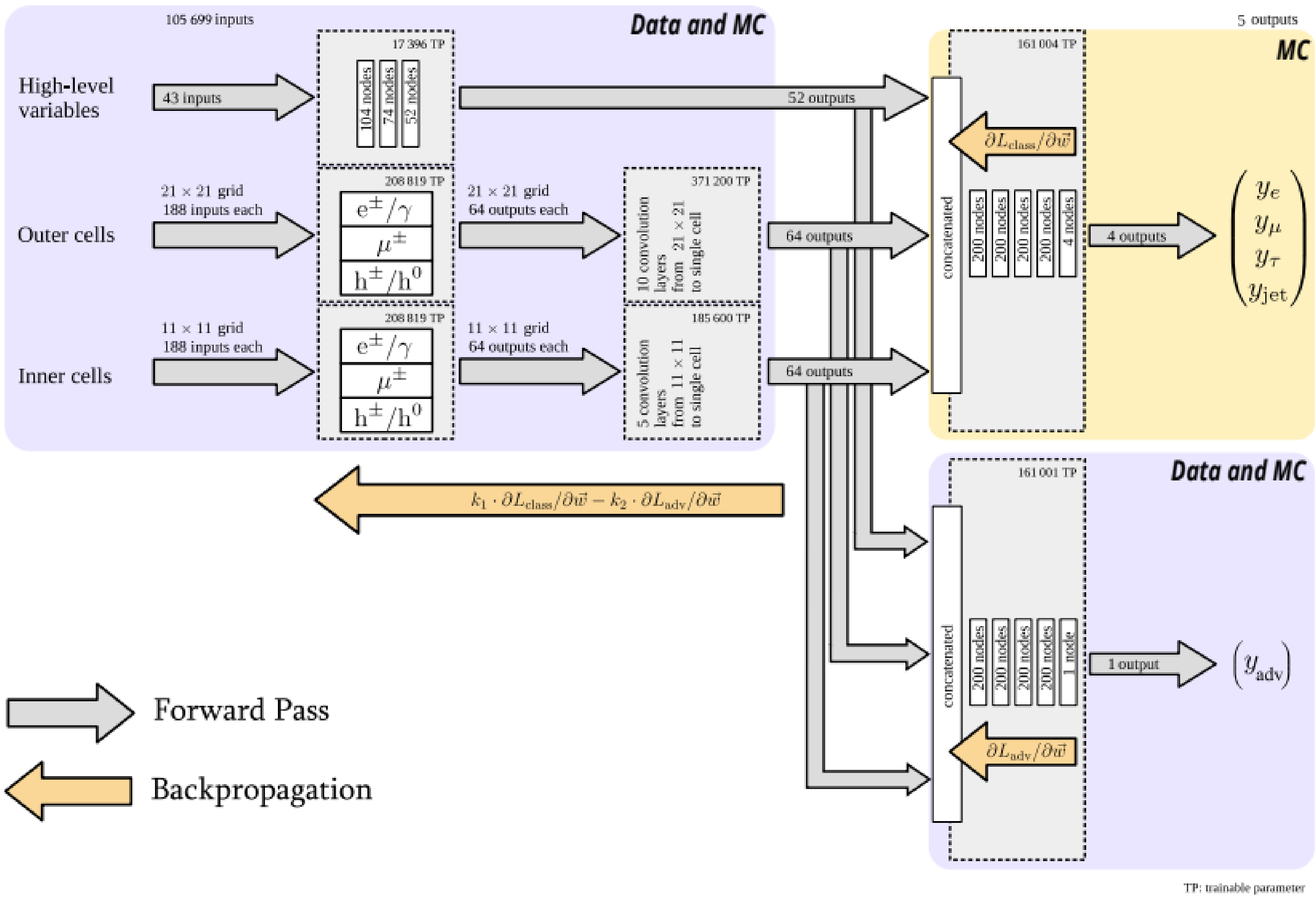
Transformer



Reconstruction: Tau ID



JINST 17 (2022) P07023

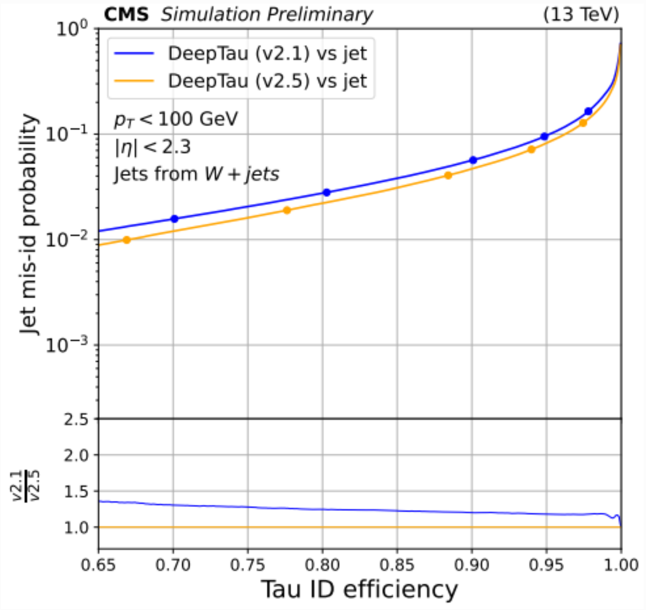


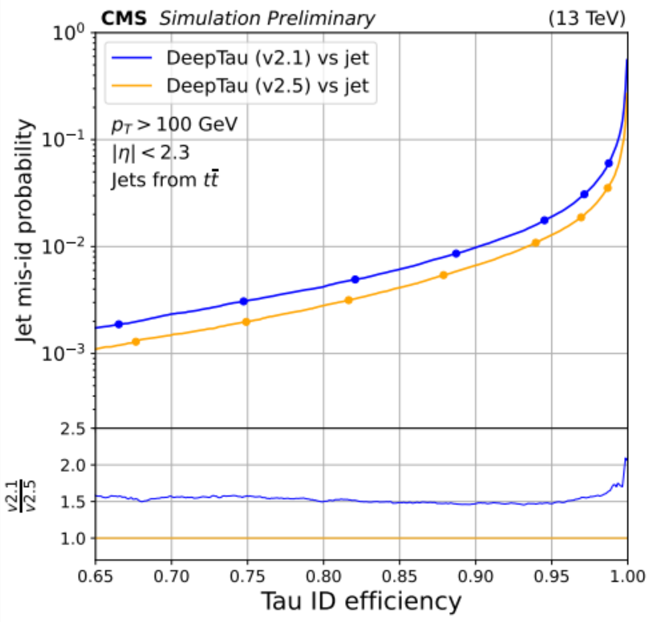
CMS DeepTau: multi-class tau identification algorithm based on CNN

- **v2.5 adds domain adaptation subnetwork with adversarial training to deal with MC mismodelling in the high-purity region**
- **Better data handling**
- **Feature standardization, hyperparameter optimization**

Reconstruction: Tau ID

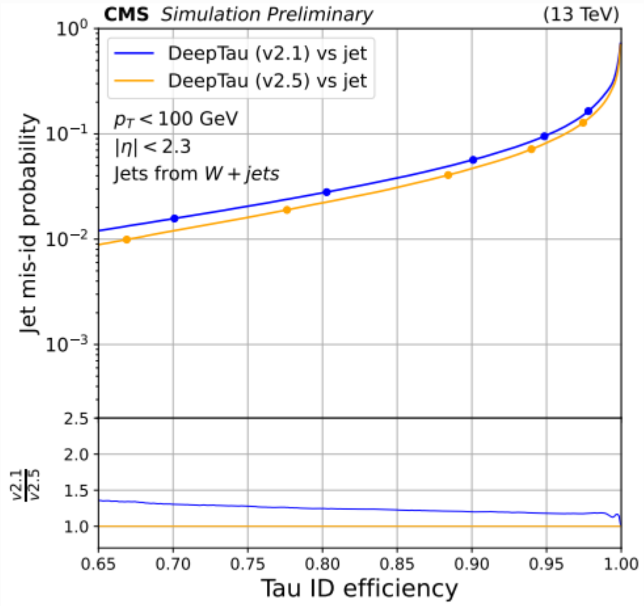
Reconstruction: Tau ID

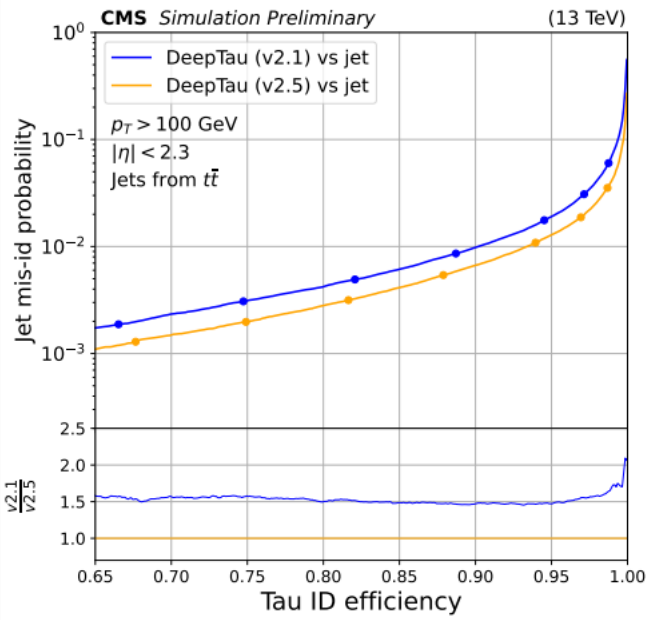


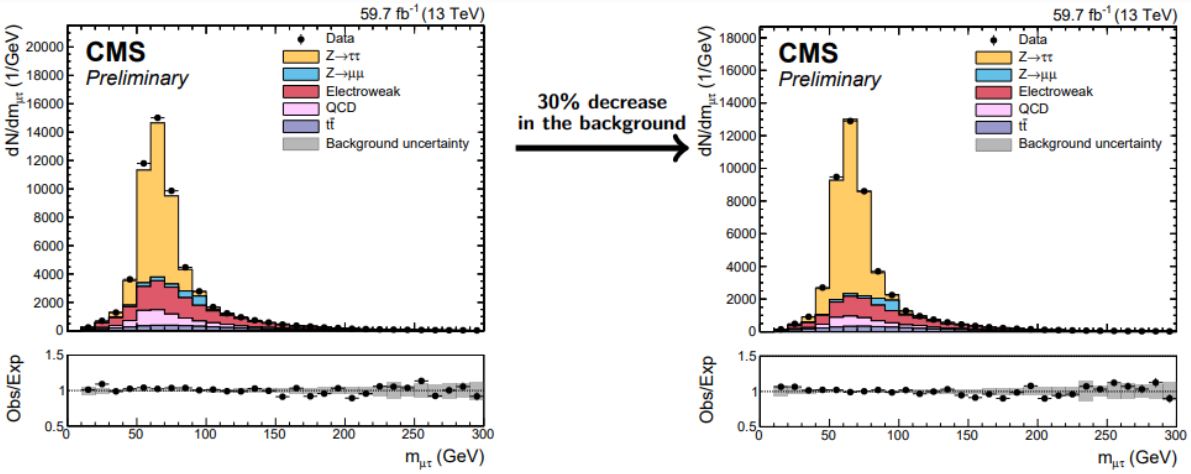


DeepTau v2.5 significant improvement compared to v2.1

- **Jet misidentification reduced by ≈ 50%**



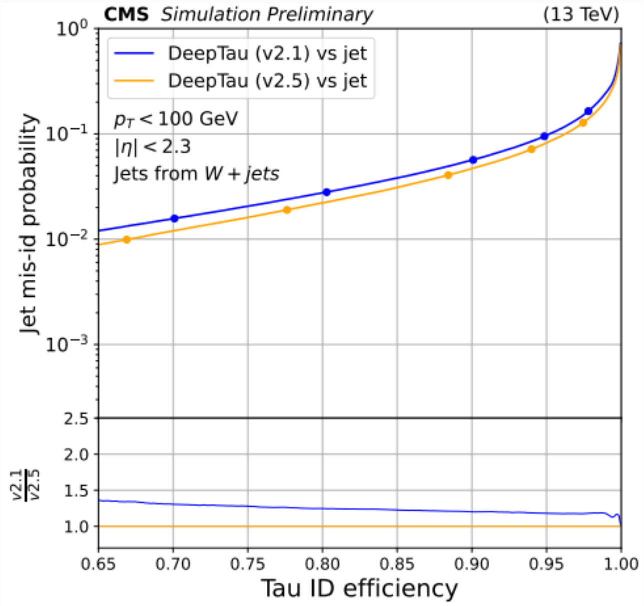


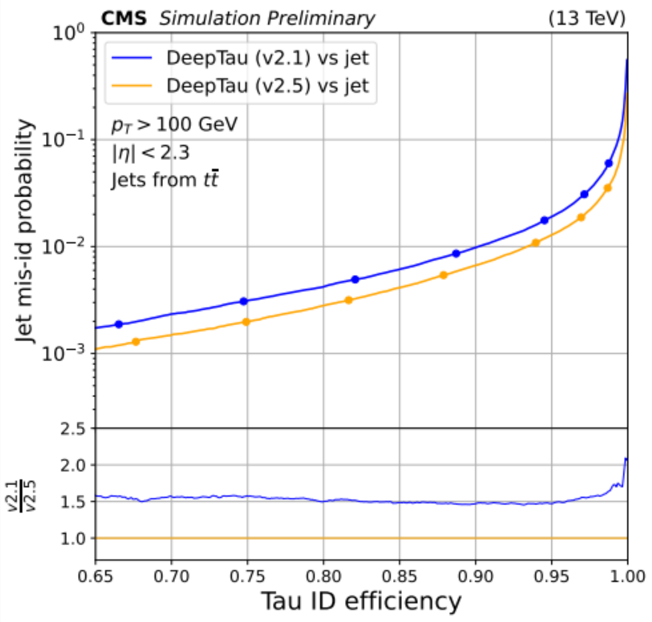


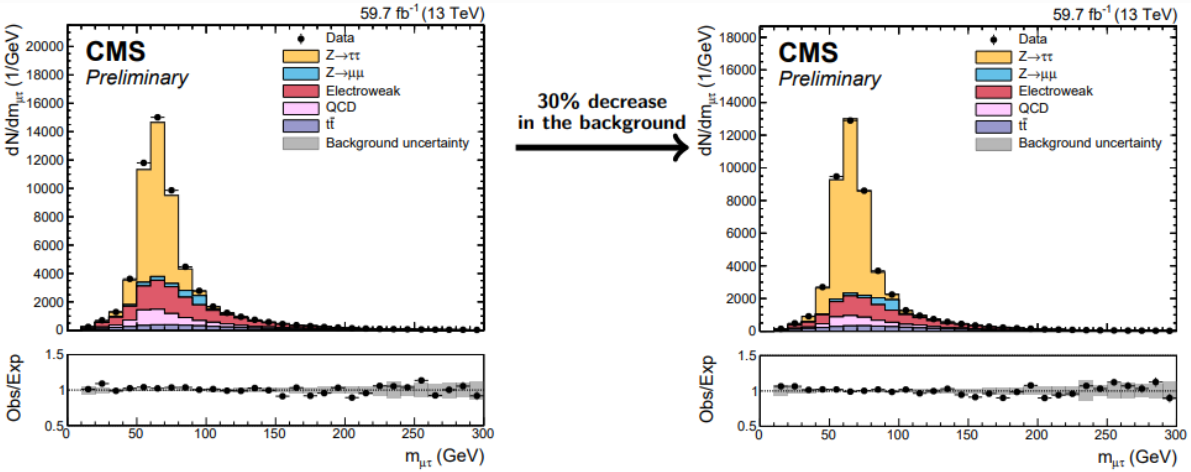
DeepTau v2.5 significant improvement compared to v2.1

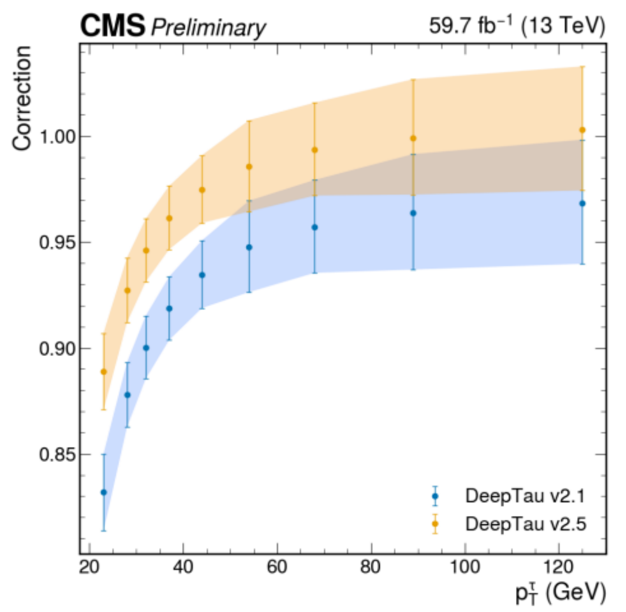
- **Jet misidentification reduced by ≈ 50%**
- **30% decrease in the background**

Reconstruction: Tau ID









[CMS-DP-2024-063](CMS-DP-2024-063)

DeepTau v2.5 significant improvement compared to v2.1

- **Jet misidentification reduced by ≈ 50%**
- **30% decrease in the background**
- **Data vs. MC scale corrections are closer to 1**
- **Minimizing dependence on MC mismodelling**

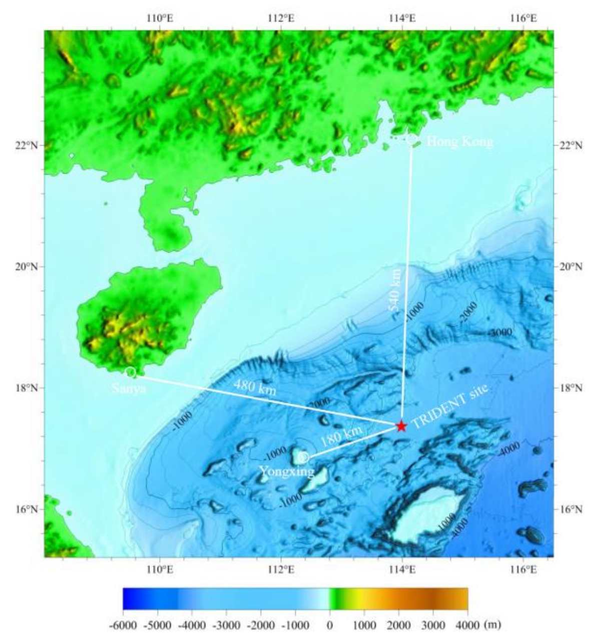
TRIDENT Experiment

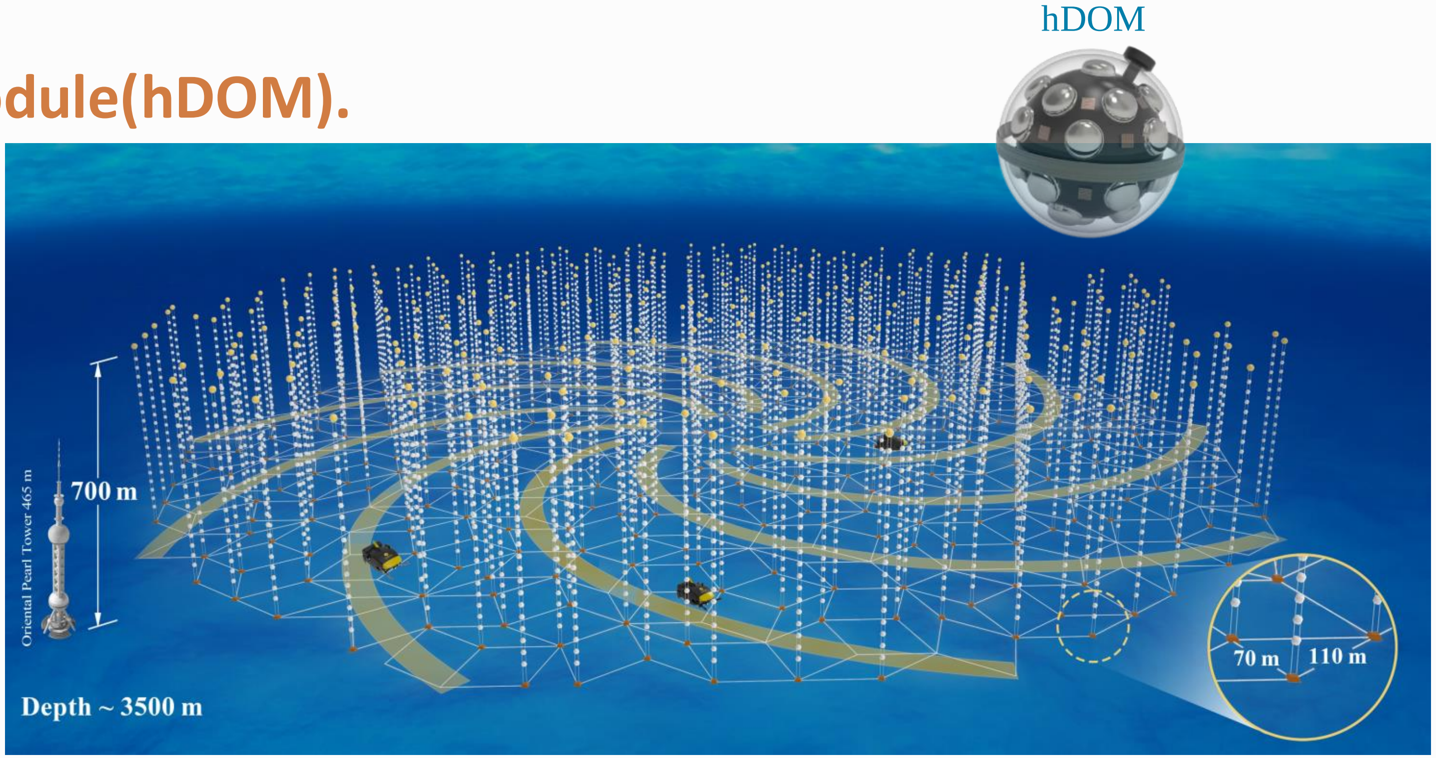
TRIDENT Experiment

- **TRIDENT: TRoplcal DEep-sea Neutrino Telescope.**

  A multi-cubic-kilometre neutrino telescope in the western Pacific Ocean. *Nature Astronomy* (2023).

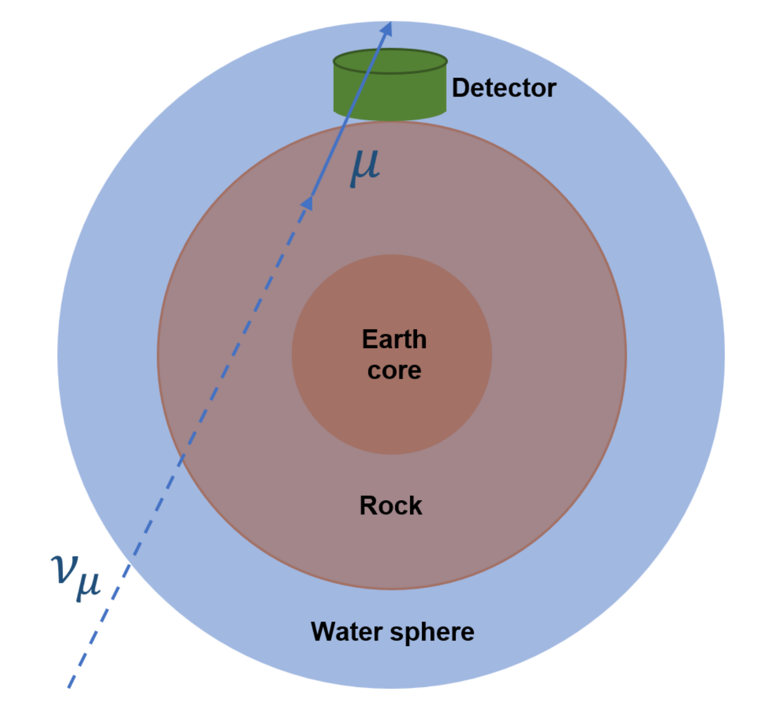
- **To be located in the South China Sea.**
- **Penrose tiling structure with 2000m radius, 700m height (8.7 km³). 3500m deep under sea level.**
- **24220 hybrid Digital Optical Module(hDOM).**





TRIDENT: Neutrino Reconstruction

TRIDENT: Neutrino Reconstruction

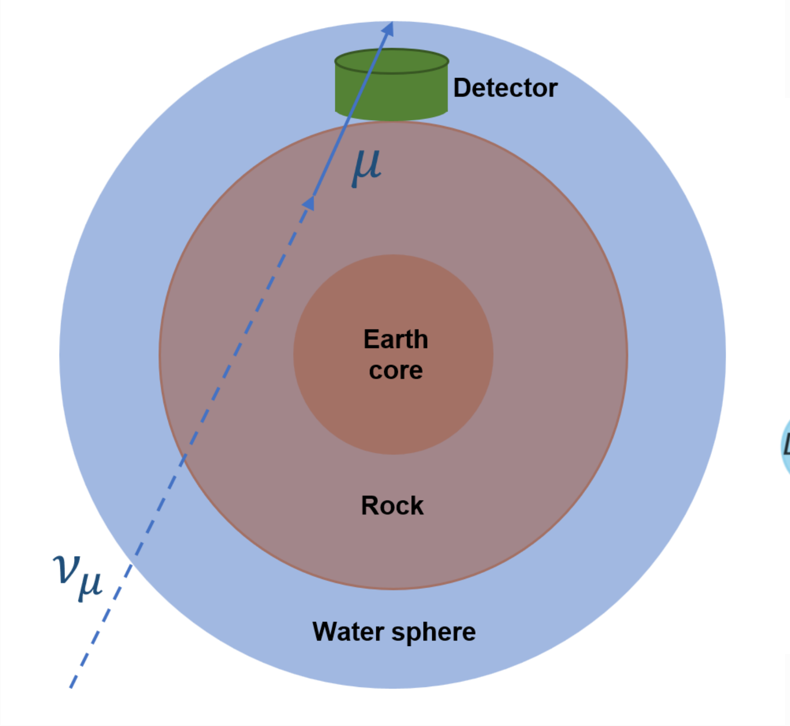


Preliminary earth model

Neutrino event generator Based on CORSIKA8

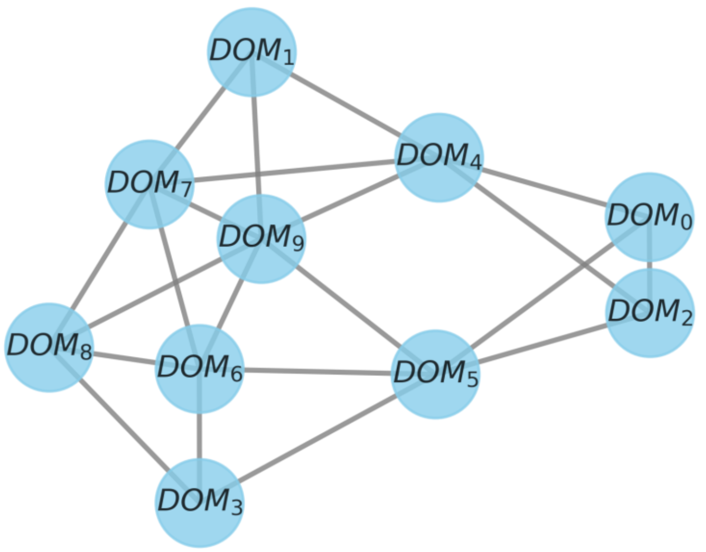
Detector simulation based on Geant4

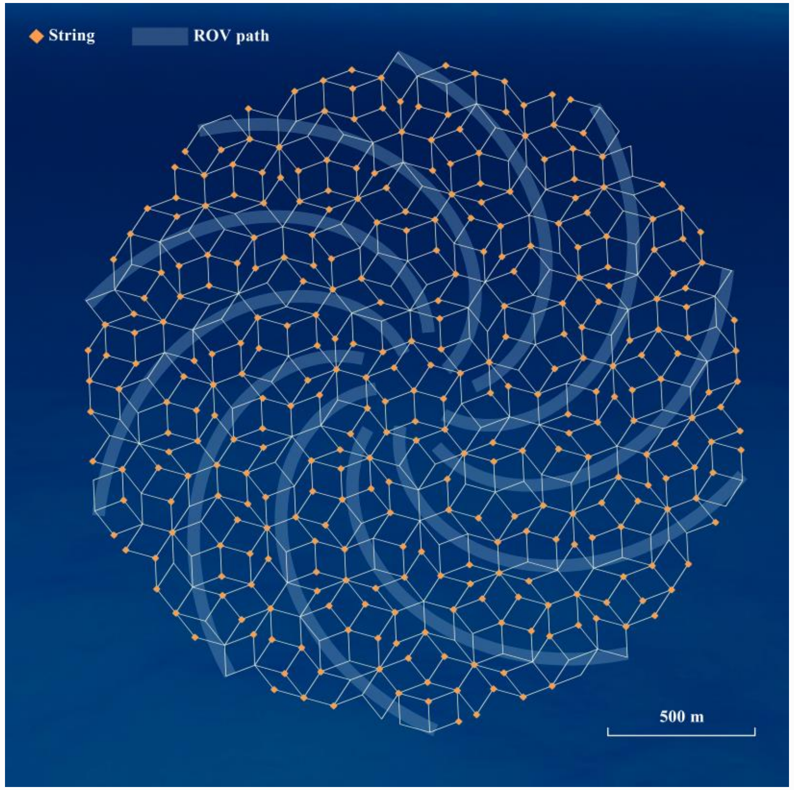
TRIDENT: Neutrino Reconstruction



Preliminary earth model

Top view of TRIDENT





Neutrino event generator Based on CORSIKA8

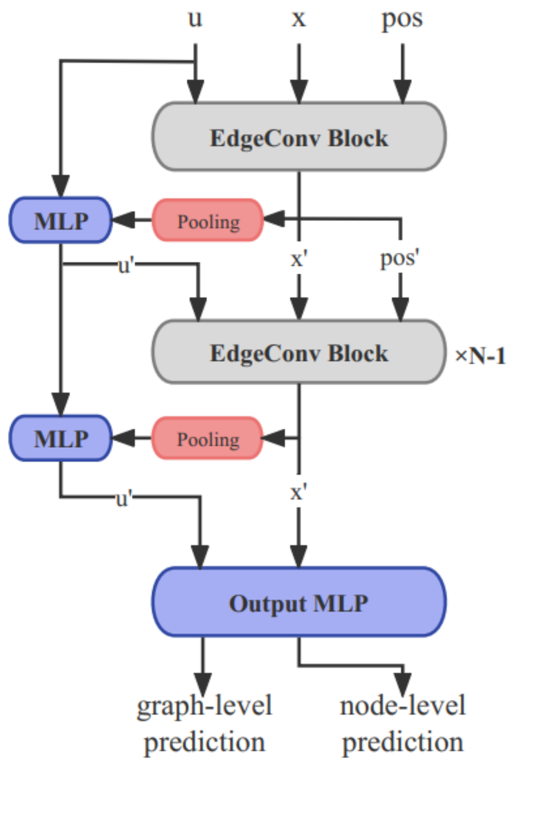
Detector simulation based on Geant4

Use point cloud to represent neutrino events:

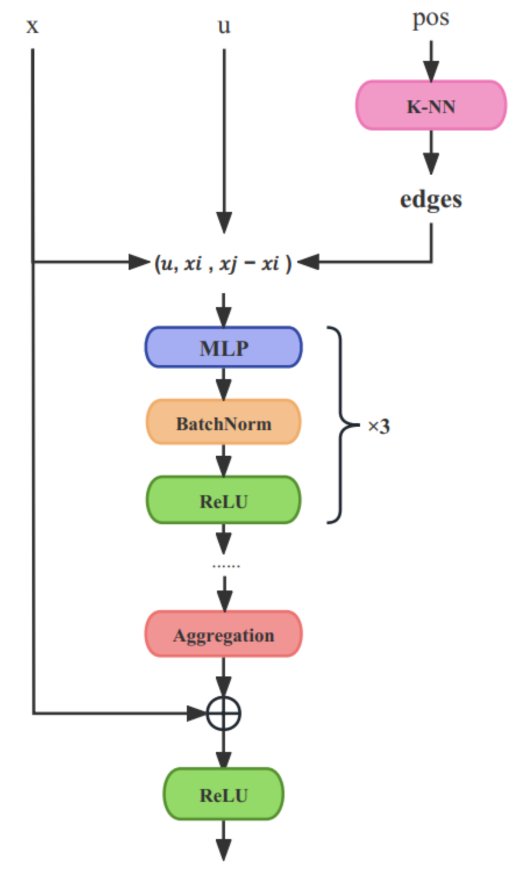
- Triggered DOMs → Nodes of point cloud
- Location of DOMs → Coordinate of nodes, pos_i.
- DOM-measured time → Features of nodes, x_i.

TRIDENT: Neutrino Reconstruction

- GNN is built based on **EdgeConv** block: modified block as in ParticleNet
- Both **graph-level** and **node-level** target can be predicted.



Network Structure

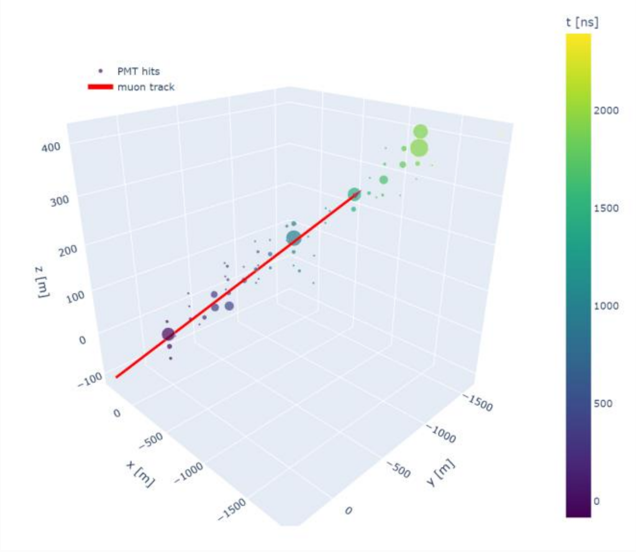


EdgeConv Block

SJTUPA 上海交通大学物理与天文学院

ν_μ Direction reconstruction

$train : validation : test = 900k : 70k : 100k$



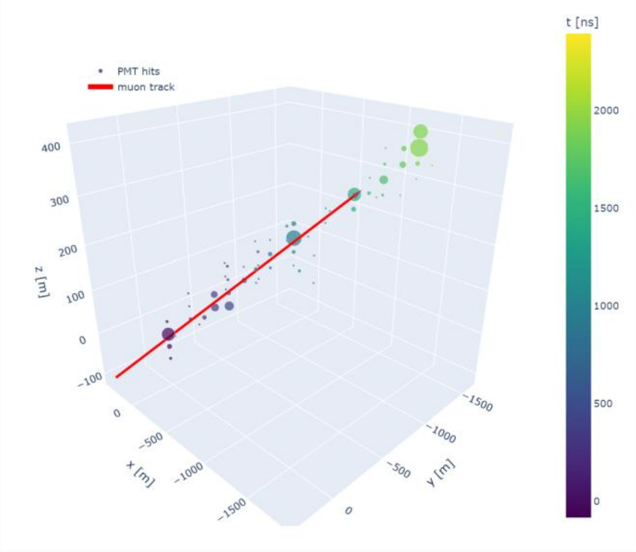
Track-like event display

SJTUPA 上海交通大学物理与天文学院

ν_μ Direction reconstruction

$train : validation : test = 900k : 70k : 100k$

- **Input features**: location $\vec{D}_i$, first photon arrival time T_i and number of photo hits n_i.

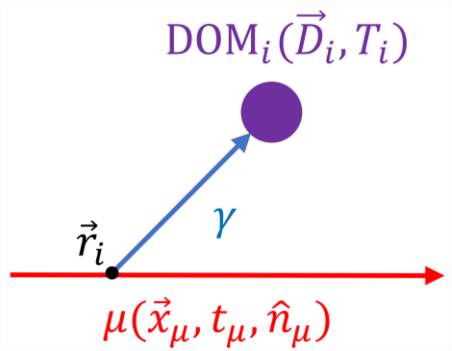


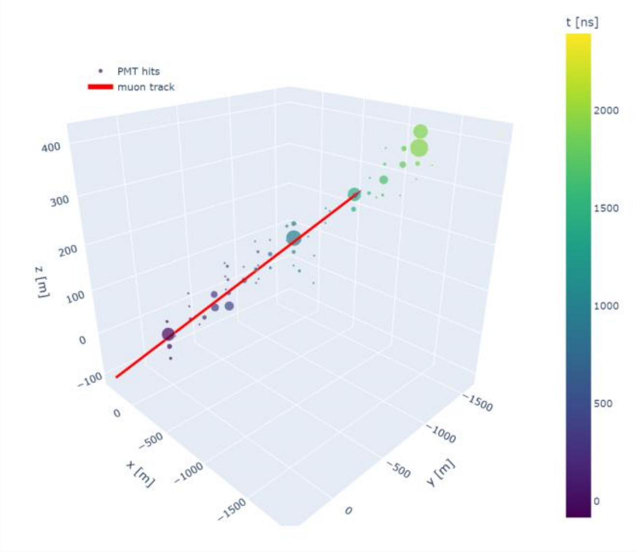
Track-like event display

ν_μ Direction reconstruction

$train : validation : test = 900k : 70k : 100k$

- **Input features**: location $\vec{D}_i$, first photon arrival time T_i and number of photo hits n_i.





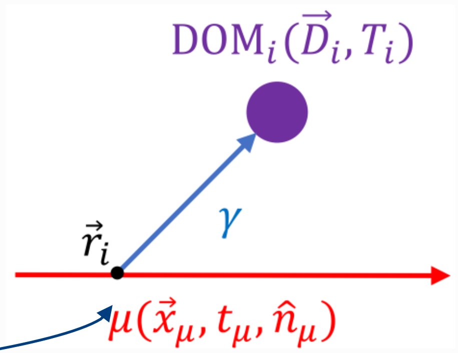
Track-like event display

ν_μ Direction reconstruction

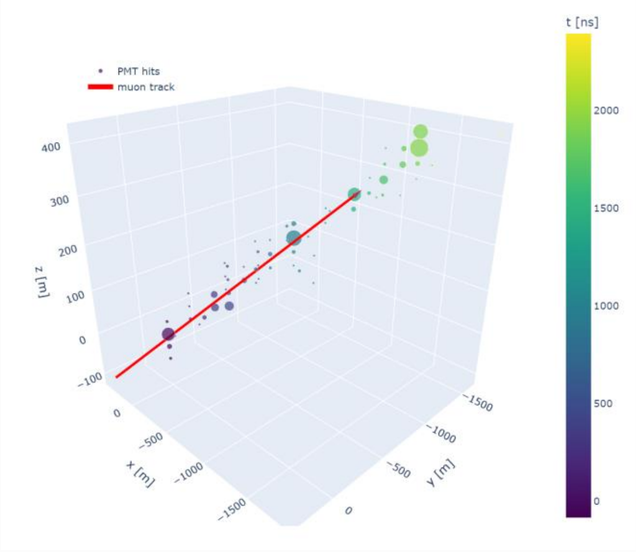
$train : validation : test = 900k : 70k : 100k$

- **Input features**: location $\vec{D}_i$, first photon arrival time T_i and number of photo hits n_i.

- To make full use of the geometric feature of track-like events, the network is trained to predict $\vec{r}_i$ for each DOM_i.



μ emits photon and triggers $\mathrm{DOM_i}$



Track-like event display

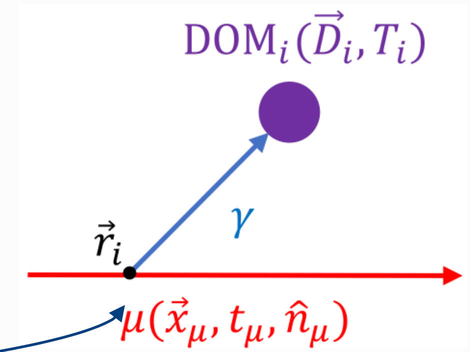
ν_μ Direction reconstruction

$train : validation : test = 900k : 70k : 100k$

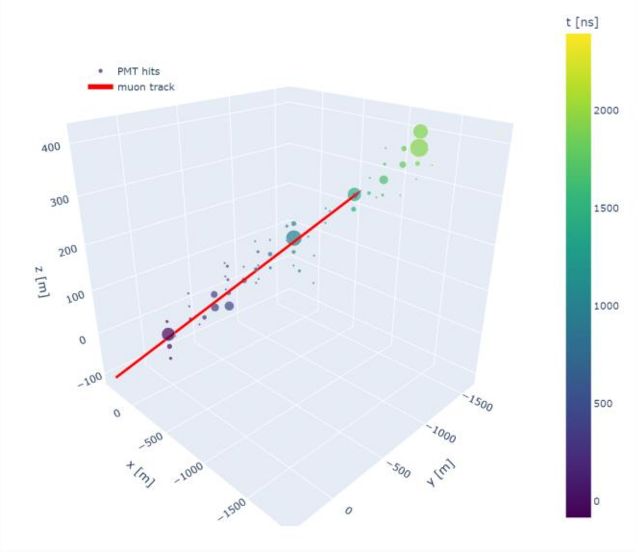
- **Input features**: location $\vec{D}_i$, first photon arrival time T_i and number of photo hits n_i.

- To make full use of the geometric feature of track-like events, the network is trained to predict $\vec{r}_i$ for each DOM_i.

- **Linear fit** on the predicted $\vec{r}'_i$ to reconstructs $\hat{n}_\mu$.



μ emits photon and triggers $\mathrm{DOM_i}$



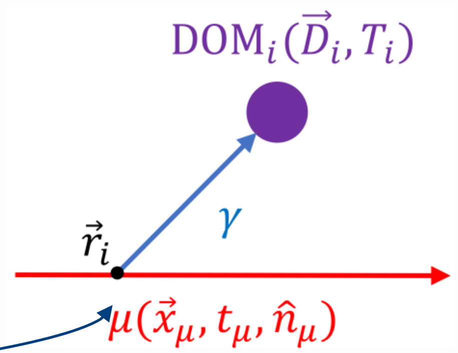
Track-like event display

ν_μ Direction reconstruction

$train : validation : test = 900k : 70k : 100k$

- **Input features**: location $\vec{D}_i$, first photon arrival time T_i and number of photo hits n_i.

- To make full use of the geometric feature of track-like events, the network is trained to predict $\vec{r}_i$ for each DOM_i.



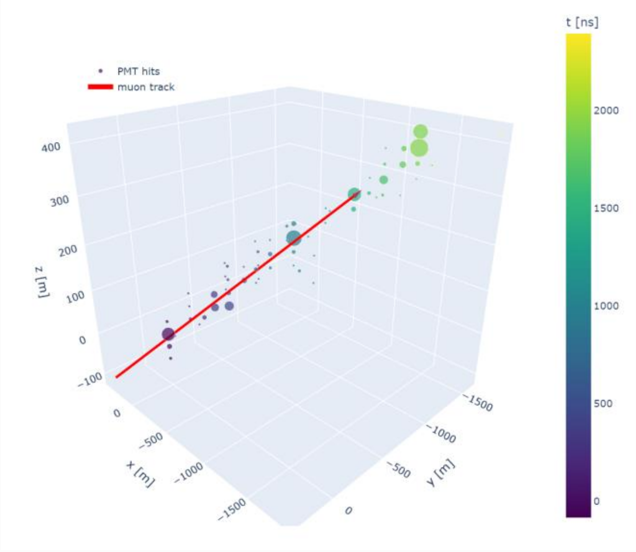
μ emits photon and triggers $\mathrm{DOM_i}$

- **Linear fit** on the predicted $\vec{r}_i'$ to reconstructs $\hat{n}_\mu$.

- **Loss function**: mean square error (MSE) with weight proportional to n_i:

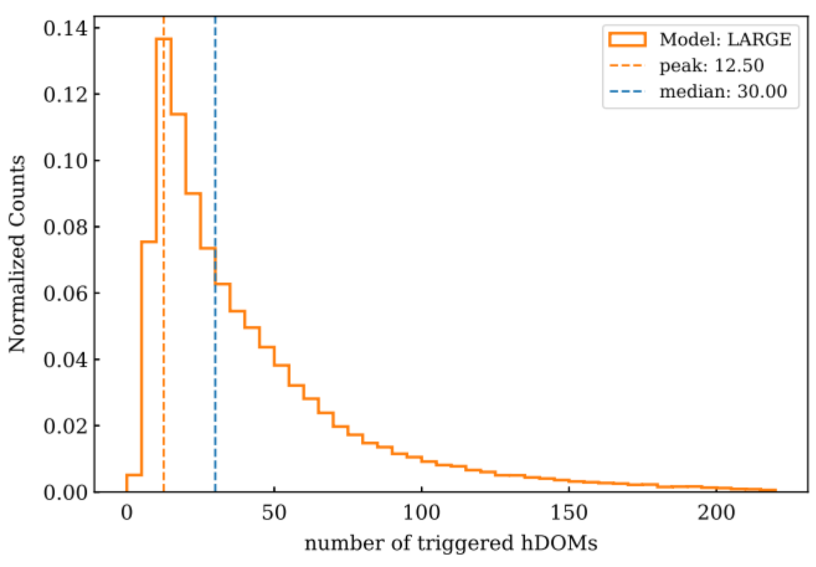
$$Loss = \Sigma_i n_i \times \left|\overrightarrow{output}_i - \vec{r}_i\right|^2 / \Sigma_i n_i$$

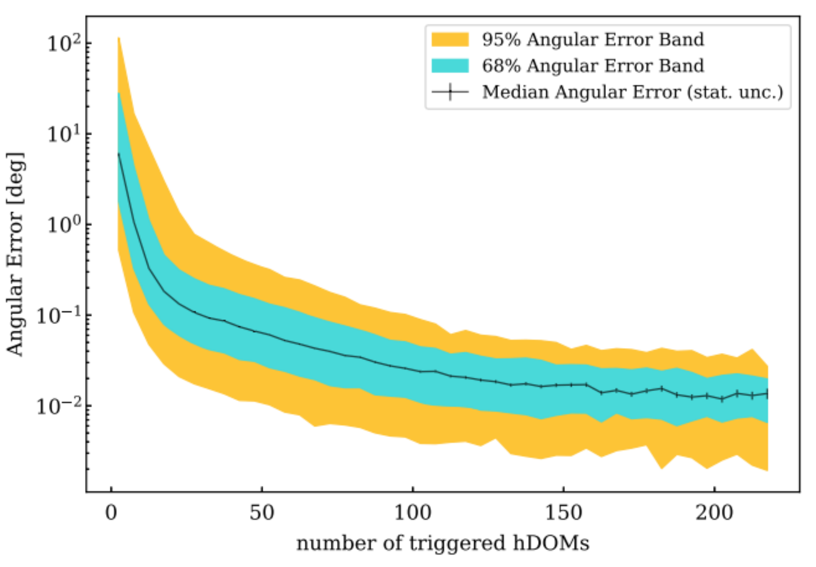
- **Hybrid-GNN models: LITE, LARGE**



Track-like event display

TRIDENT: Neutrino Reconstruction





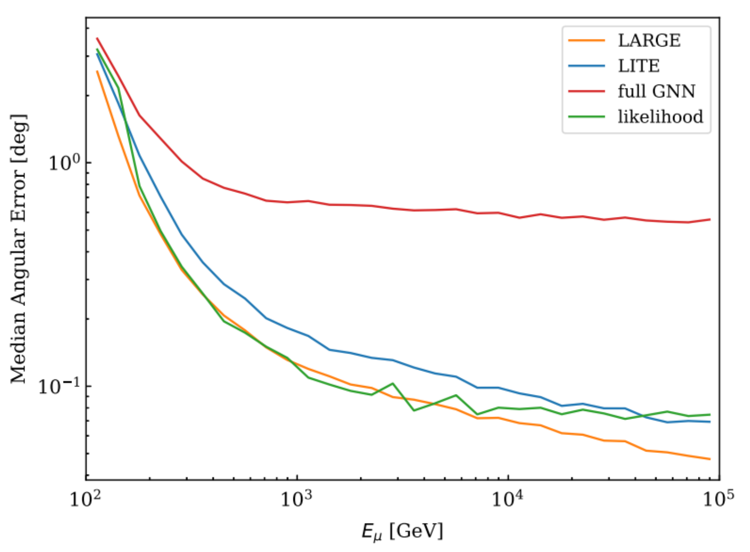


TABLE I. Mean run-time cost per inference.

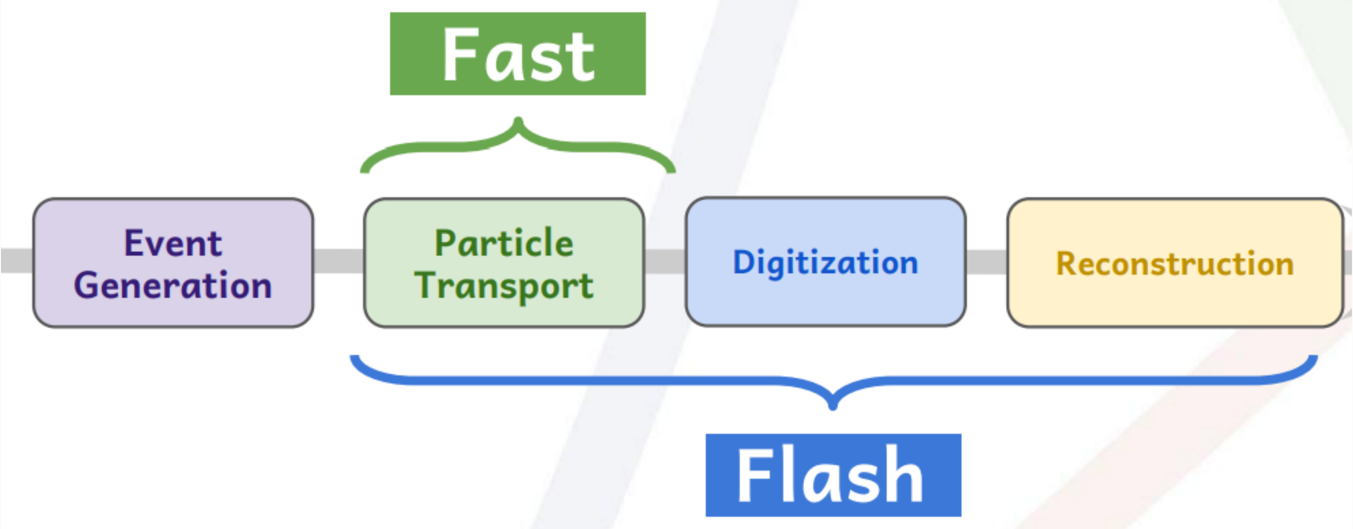
Method	Time (0.1–1 TeV) (ms)	Time (1–10 TeV) (ms)	Time (10–100 TeV) (ms)
Likelihood	1552.30	1259.86	919.14
GNN light (GPU)	0.19	0.21	0.29
GNN large (GPU)	0.38	0.78	2.37
GNN light (CPU)	5.05	12.53	30.44
GNN large (CPU)	54.71	152.48	181.80

Phys. Rev. D 112, 072012

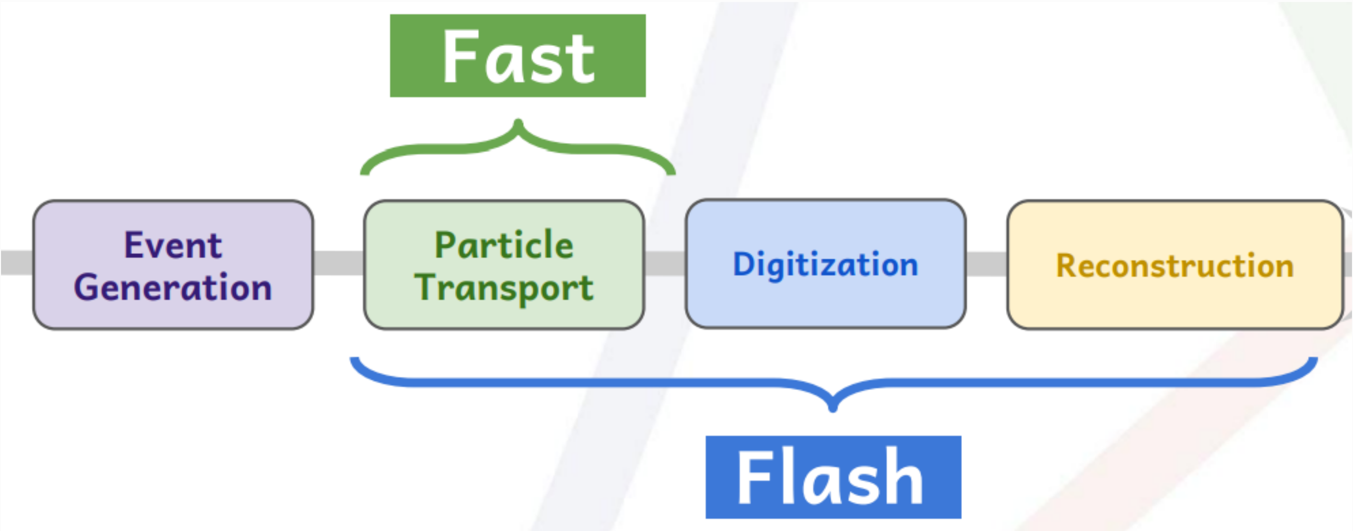
- **Median angular error decreases from 1 degree to 0.1 degree as the energy of ν_μ increases**
- **Light hybrid-GNN model (LITE) runs 0.19–0.29 ms per event on GPUs, 1000 times faster than traditional likelihood fitting method --- real time processing**
- **Large hybrid-GNN model (LARGE) takes longer but with more precision --- offline processing**

Fast and Flash Simulations at LHCb

Fast and Flash Simulations at LHCb

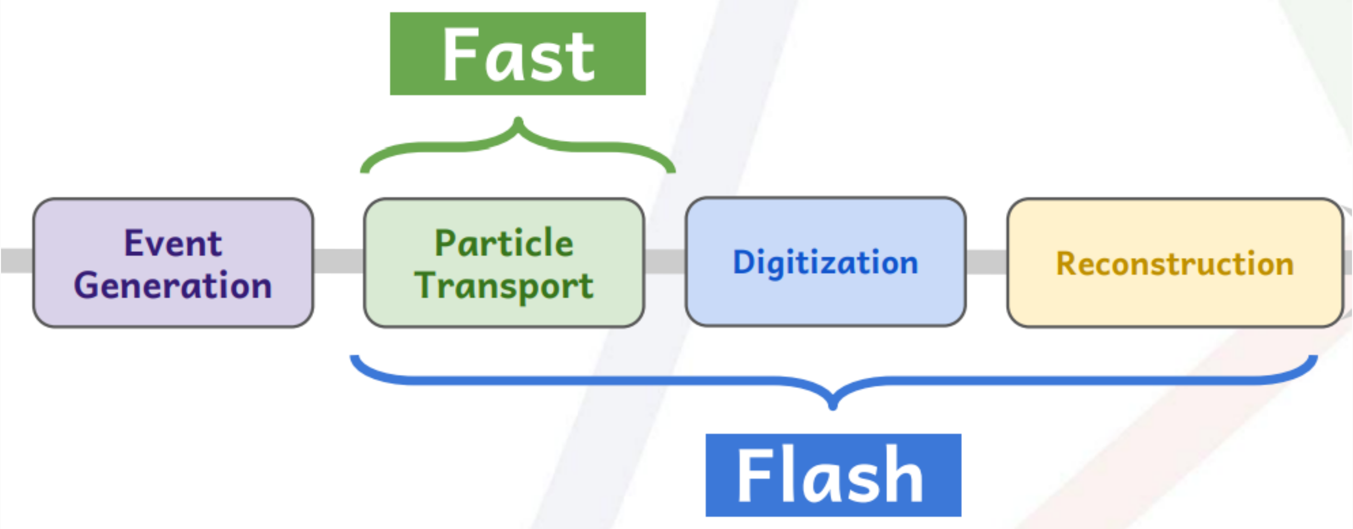


Fast and Flash Simulations at LHCb



Why simulation matters?

Fast and Flash Simulations at LHCb

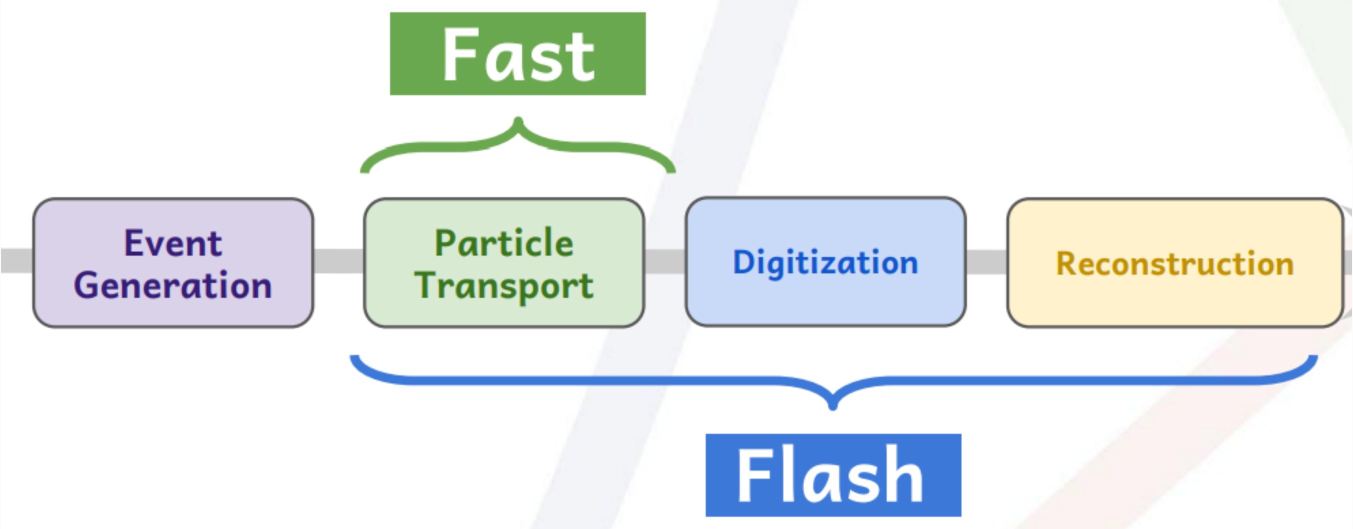


Why simulation matters?

- **90% of computing resources are used for simulations at LHCb**
- **Calorimeter simulation is the most computationally intensive part of the simulation process**
- **60% of the total CPU time is used for calorimeter simulations**

Fast and Flash Simulations at LHCb

arXiv:2511.02020



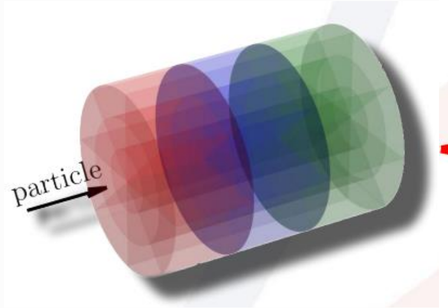
Why simulation matters?

- 90% of computing resources are used for simulations at LHCb
- Calorimeter simulation is the most computationally intensive part of the simulation process
- 60% of the total CPU time is used for calorimeter simulations

CaloML based on CaloChallenge

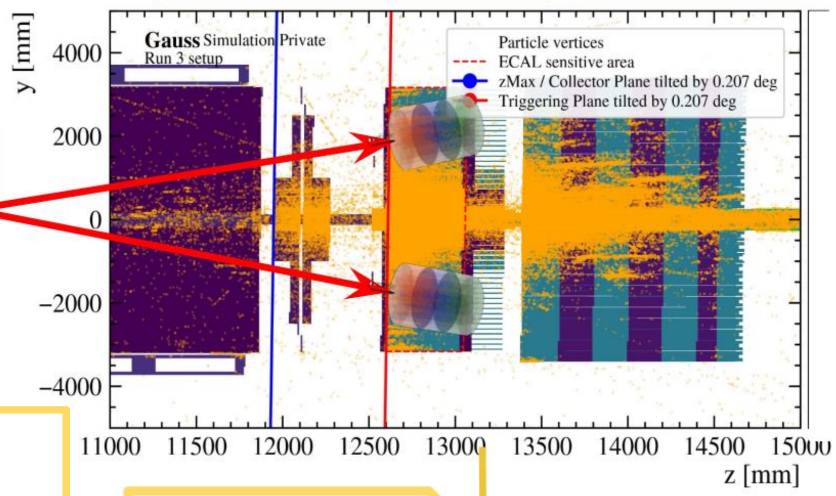
- CaloML is the first production-ready option with generative models

Fast Simulation at LHCb



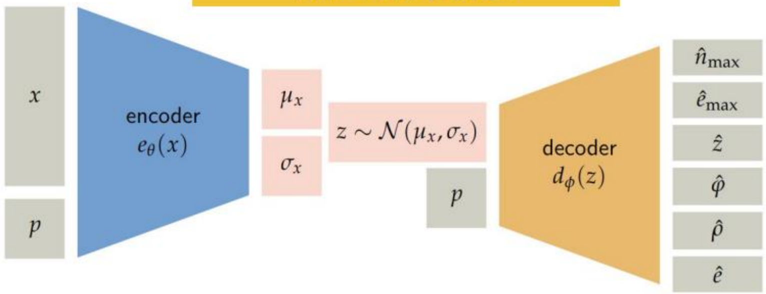
Cylinder of virtual hits around a particle shower

LHCb Training Data



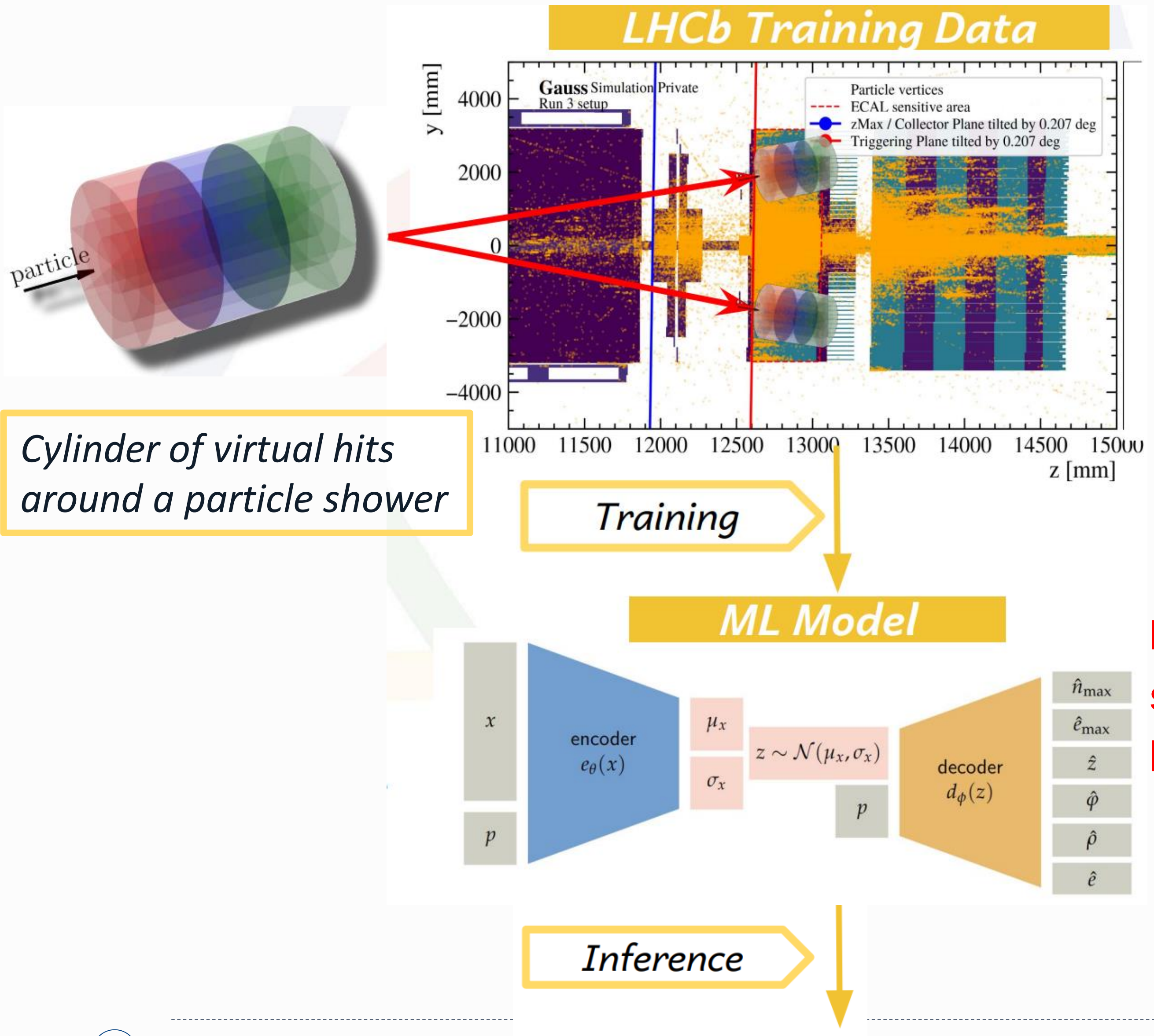
Training

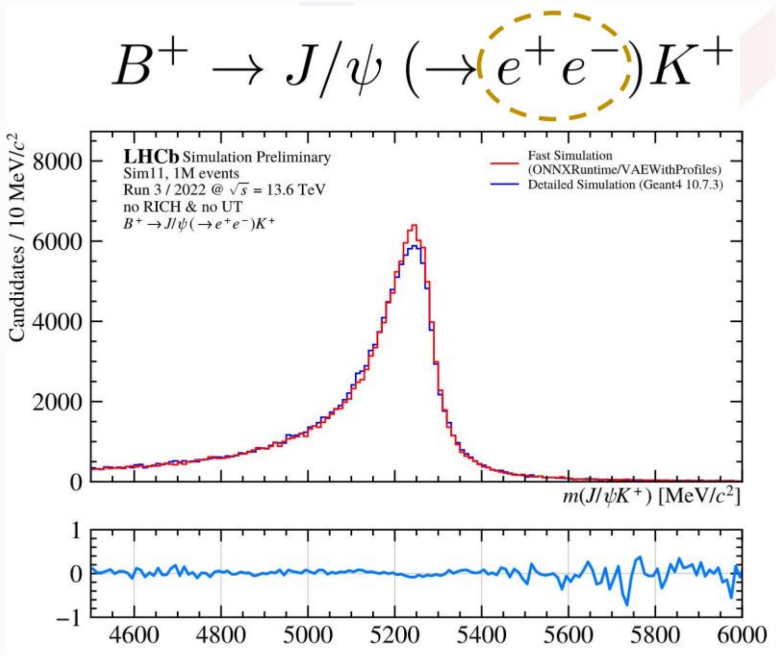
ML Model



Inference

Fast Simulation at LHCb



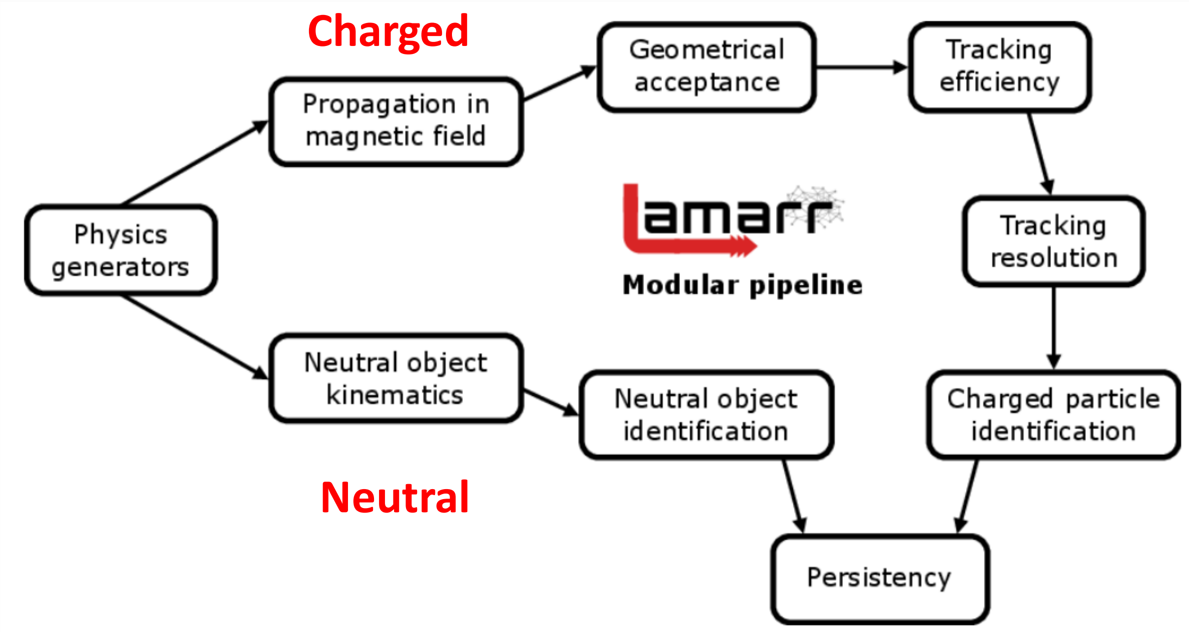


Modified Variational autoencoders (VAE) predict spatial and energy profiles of the cylinders, improving both accuracy and training speed

- **~100x times faster for electrons and photons in ECAL**
- **~0.01% energy difference on reconstructed objects**
- **Ongoing efforts to include hadrons**
- **Good agreement with physics observables**

Flash Simulation at LHCb

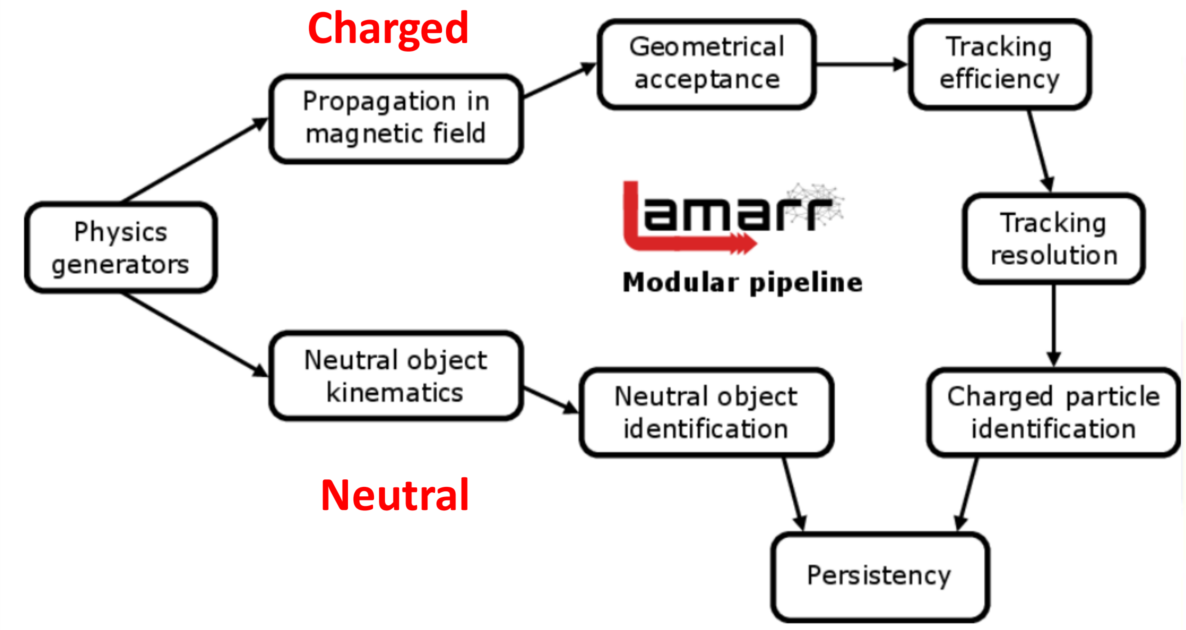
Flash Simulation at LHCb

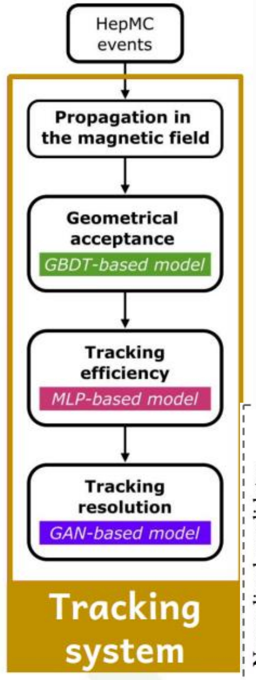


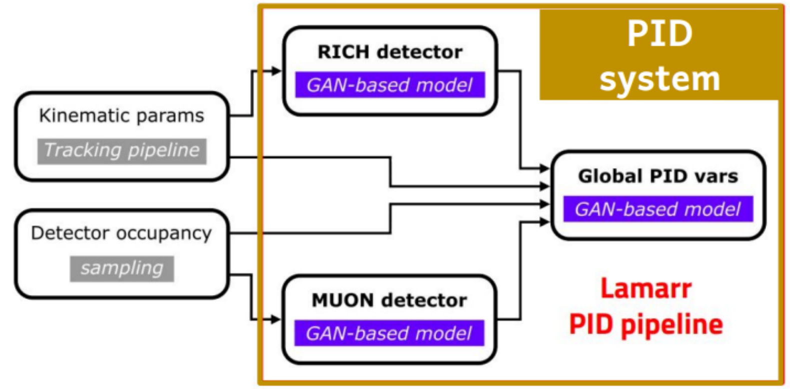
Two branches approach

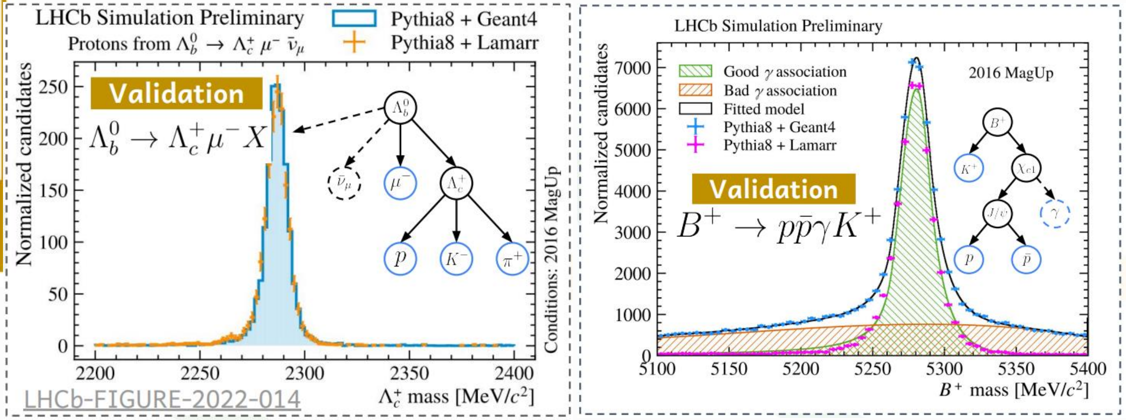
- **Charged**: branch treating charged particles relying on tracking and particle identification parameterizations
- **Neutral**: branch treating neutral particles that require an accurate parameterization of the calorimeter

Flash Simulation at LHCb







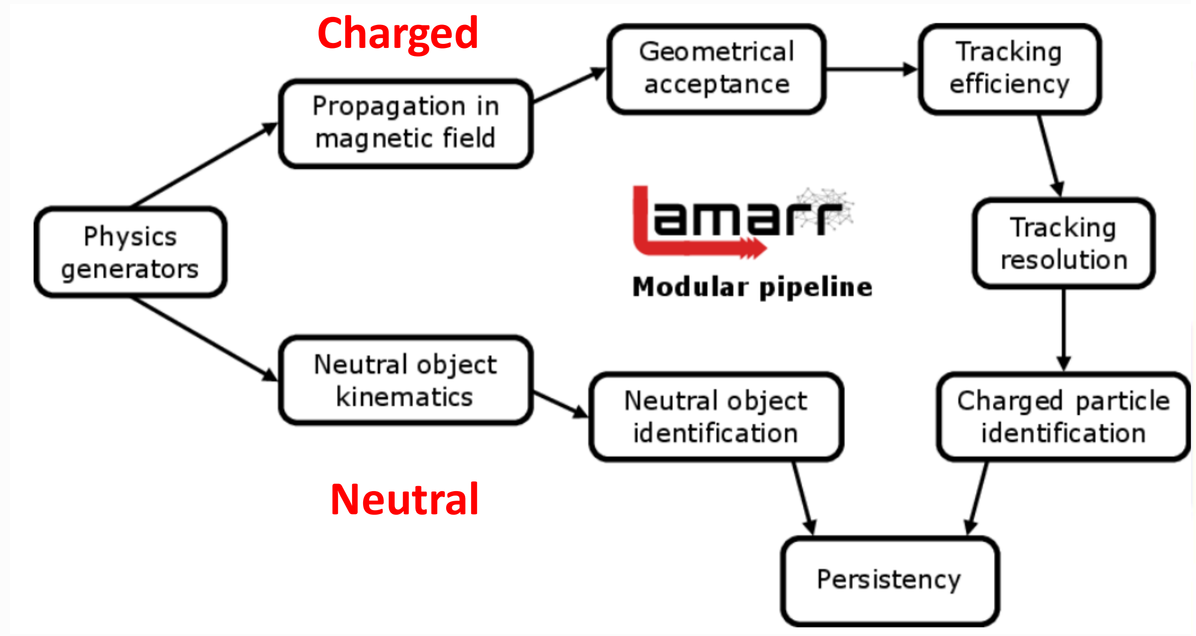


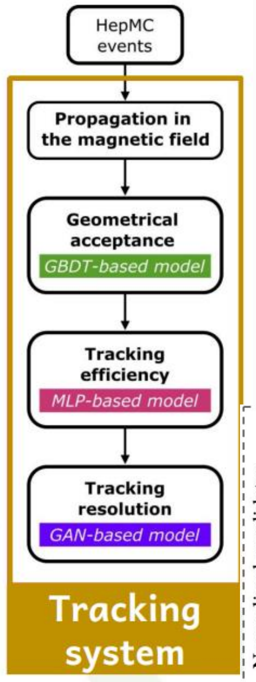
Two branches approach

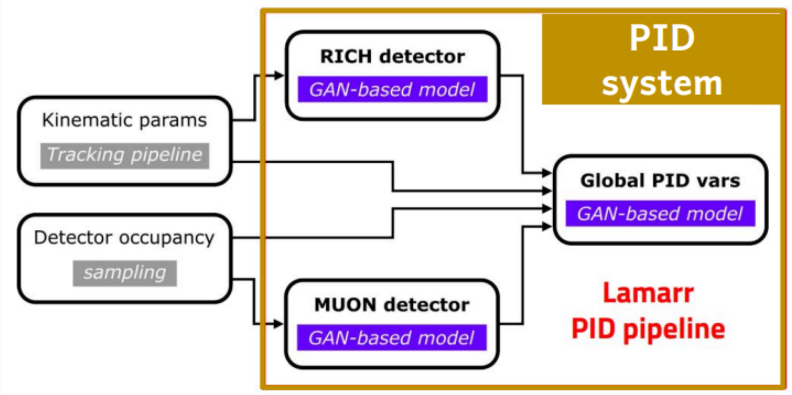
- **Charged**: branch treating charged particles relying on tracking and particle identification parameterizations
- **Neutral**: branch treating neutral particles that require an accurate parameterization of the calorimeter

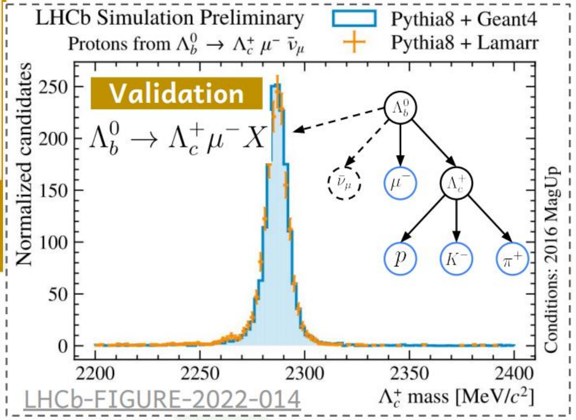
Flash Simulation at LHCb

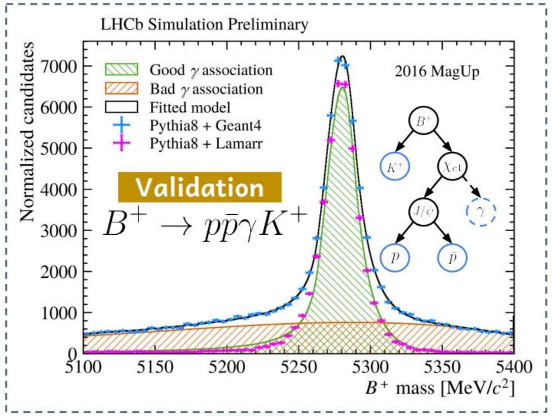
SJTUPA 上海交通大学物理与天文学院











Two branches approach

- **Charged**: branch treating charged particles relying on tracking and particle identification parameterizations
- **Neutral**: branch treating neutral particles that require an accurate parameterization of the calorimeter

- **Lamarr** accelerates detector simulation and reconstruction by 2-3 orders of magnitude compared to GEANT4 full simulation
- **Validation** for LHCb analyses, neutral sector needs more work

Large Language Models for Design

Motivation

• Physics instruments (e.g., collider detectors) require long, expensive design cycles.
• ML optimization exists (Trust-region (TR) optimizer, differentiable surrogates, RL), but humans still craft action spaces, rewards, and workflows.

Can LLMs propose physically meaningful designs with *only prompting*?

- Keep simulator + reward fixed, swap proposal mechanism (e.g. RL → LLM prompting).

Large Language Models for Design

arXiv:2601.07580

Motivation

- Physics instruments (e.g., collider detectors) require long, expensive design cycles.
- ML optimization exists (Trust-region (TR) optimizer, differentiable surrogates, RL), but humans still craft action spaces, rewards, and workflows.

Can LLMs propose physically meaningful designs with *only prompting*?

- Keep simulator + reward fixed, swap proposal mechanism (e.g. RL → LLM prompting).

Benchmarks reused from RL study (controlled testbeds)

A) Sampling calorimeter segmentation

Design variables
- layer positions z (mm)
- discrete layer thickness t
- global thickness/cost budget

Metric
Mean-corrected energy resolution
(EM & hadronic @ 50/100 GeV)

B) Magnetic spectrometer layout

Design variables
- station positions z (m)
- granularity g (bins/side)
- total pixel budget

Metric
Tracking efficiency & momentum resolution @ 10/100 GeV

Large Language Models for Design

Large Language Models for Design

Prompt

Problem spec
+ constraints
+ objective targets
+ memory: best designs

Large Language Models for Design

Prompt

Problem spec
+ constraints
+ objective targets
+ memory: best designs

→

LLM proposes

Return ONLY JSON
{z:[...], t:[...]}
or {z:[...], g:[...]}

Large Language Models for Design

Prompt

Problem spec
+ constraints
+ objective targets
+ memory: best designs

→

LLM proposes

Return ONLY JSON
{z:[...], t:[...]}
or {z:[...], g:[...]}

→

Projection

Sort & snap to discrete sets
enforce budgets
remove overlaps

Large Language Models for Design

Prompt

Problem spec
+ constraints
+ objective targets
+ memory: best designs

→

LLM proposes

Return ONLY JSON
{z:[...], t:[...]}
or {z:[...], g:[...]}

→

Projection

Sort & snap to discrete sets
enforce budgets
remove overlaps

→

Evaluate / Iterate

Simulator + reconstruction
+ reward S(x)

Large Language Models for Design

Prompt

Problem spec
+ constraints
+ objective targets
+ memory: best designs

→

LLM proposes

Return ONLY JSON
{z:[...], t:[...]}
or {z:[...], g:[...]}

→

Projection

Sort & snap to discrete sets
enforce budgets
remove overlaps

→

Evaluate / Iterate

Simulator + reconstruction
+ reward S(x)

Optional hybrid step: Trust-Region (TR) refinement

- Keep discrete choices fixed
- Locally optimize continuous positions (z)
- Use black-box optimizer (BOBYQA) under hard constraints

Large Language Models for Design

Prompt
Problem spec
+ constraints
+ objective targets
+ memory: best designs

→

LLM proposes
Return ONLY JSON
{z:[...], t:[...]}
or {z:[...], g:[...]}

→

Projection
Sort & snap to discrete sets
enforce budgets
remove overlaps

→

Evaluate / Iterate
Simulator + reconstruction
+ reward S(x)

What makes it interesting?

- No fine-tuning, no gradients, no simulator interaction by the model.
- LLM is used as a *proposal generator* using broad pretrained physics knowledge.
- Feasibility projection prevents wasting evaluations on invalid designs.
- Memory of best designs gives a compact "dataset" for in-context improvement.

Optional hybrid step: Trust-Region (TR) refinement

- Keep discrete choices fixed
- Locally optimize continuous positions (z)
- Use black-box optimizer (BOBYQA) under hard constraints

Large Language Models for Design

SJTUPA 上海交通大学物理与天文学院

Models tested (350 proposal iterations each)

GPT-OSS-20B • GPT-OSS-120B • GPT-5 • Gemini 2.5 Pro

Calorimeter benchmark

Highlight (hadronic resolution dominates reward)

Baseline	Had 50 GeV: 32.13%	Had 100 GeV: 25.19%
RL best	Had 50 GeV: 24.29%	Had 100 GeV: 18.07%
Best LLM(+TR)	Had 50 GeV: 25.09%	Had 100 GeV: 18.06%

Observation: even without task-specific training, LLMs quickly find non-uniform layer layouts that improve hadronic performance.

Spectrometer benchmark

Highlight (100 GeV momentum resolution)

Baseline	Res@100 GeV: 13.27%	Eff@100 GeV: 99.17%
RL best	Res@100 GeV: 7.95%	Eff@100 GeV: 99.90%
Best LLM(+TR)	Res@100 GeV: 7.97%	Eff@100 GeV: 99.91%

Observation: open-weight GPT-OSS-20B performs strongly; TR improves z-placement and nearly matches RL at 100 GeV.

Large Language Models for Design

SJTUPA 上海交通大学物理与天文学院

Main takeaways

LLMs can generate valid designs under hard constraints
Even with no task-specific training, prompting + memory yields physically meaningful layouts.

RL remains the strongest end-to-end optimizer
But LLM+local refinement can recover much of the performance.

Feasibility projection is crucial
Deterministic cleanup turns brittle generations into a stable search process.

LLMs as meta-planners
They can help define search strategies, organize experiments, and orchestrate optimization pipelines.

Large Language Models for Design

Main takeaways

LLMs can generate valid designs under hard constraints
Even with no task-specific training, prompting + memory yields physically meaningful layouts.

RL remains the strongest end-to-end optimizer
But LLM+local refinement can recover much of the performance.

Feasibility projection is crucial
Deterministic cleanup turns brittle generations into a stable search process.

LLMs as meta-planners
They can help define search strategies, organize experiments, and orchestrate optimization pipelines.

A practical hybrid workflow (toward "closed-loop" design)

LLM
Propose design hypotheses, constraints, and evaluation plan

Optimization engine
RL / TR / differentiable surrogate refines designs under reward

Simulation & validation
GEANT4-like simulation, reconstruction, system-level checks

Human-in-the-loop
Review, constraint updates, safety & engineering feasibility

Large Language Models for Design

Main takeaways

LLMs can generate valid designs under hard constraints
Even with no task-specific training, prompting + memory yields physically meaningful layouts.

RL remains the strongest end-to-end optimizer
But LLM+local refinement can recover much of the performance.

Feasibility projection is crucial
Deterministic cleanup turns brittle generations into a stable search process.

LLMs as meta-planners
They can help define search strategies, organize experiments, and orchestrate optimization pipelines.

Limitations & Outlook

- Benchmarks are simplified; real detector design adds more subsystems and constraints.
- LLMs need robust guardrails (projection, validation) to avoid invalid or misleading proposals.
- Agent that calls LLMs and tools to do optimization studies.

A practical hybrid workflow (toward "closed-loop" design)

LLM
Propose design hypotheses, constraints, and evaluation plan

Optimization engine
RL / TR / differentiable surrogate refines designs under reward

Simulation & validation
GEANT4-like simulation, reconstruction, system-level checks

Human-in-the-loop
Review, constraint updates, safety & engineering feasibility

Model-centric vs. Agent-centric

Model-centric vs. Agent-centric

Model-centric AI

Model-centric vs. Agent-centric

SJTUPA 上海交通大学物理与天文学院

Model-centric AI

- Single inference

Model-centric vs. Agent-centric

Model-centric AI

- Single inference
- Passive response

Model-centric vs. Agent-centric

SJTUPA 上海交通大学物理与天文学院

Model-centric AI

- Single inference
- Passive response
- Stateless

Model-centric vs. Agent-centric

SJTUPA 上海交通大学物理与天文学院

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives

Model-centric vs. Agent-centric

SJTUPA 上海交通大学物理与天文学院

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Model-centric vs. Agent-centric

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Agent-centric AI

Model-centric vs. Agent-centric

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Agent-centric AI

- Long-horizon with multiple tasks

Model-centric vs. Agent-centric

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Agent-centric AI

- Long-horizon with multiple tasks
- Proactive planning

Model-centric vs. Agent-centric

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Agent-centric AI

- Long-horizon with multiple tasks
- Proactive planning
- State memorized

Model-centric vs. Agent-centric

SJTUPA 上海交通大学物理与天文学院

Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Agent-centric AI

- Long-horizon with multiple tasks
- Proactive planning
- State memorized
- Can understand goals and objectives

Model-centric vs. Agent-centric

SJTUPA 上海交通大学物理与天文学院

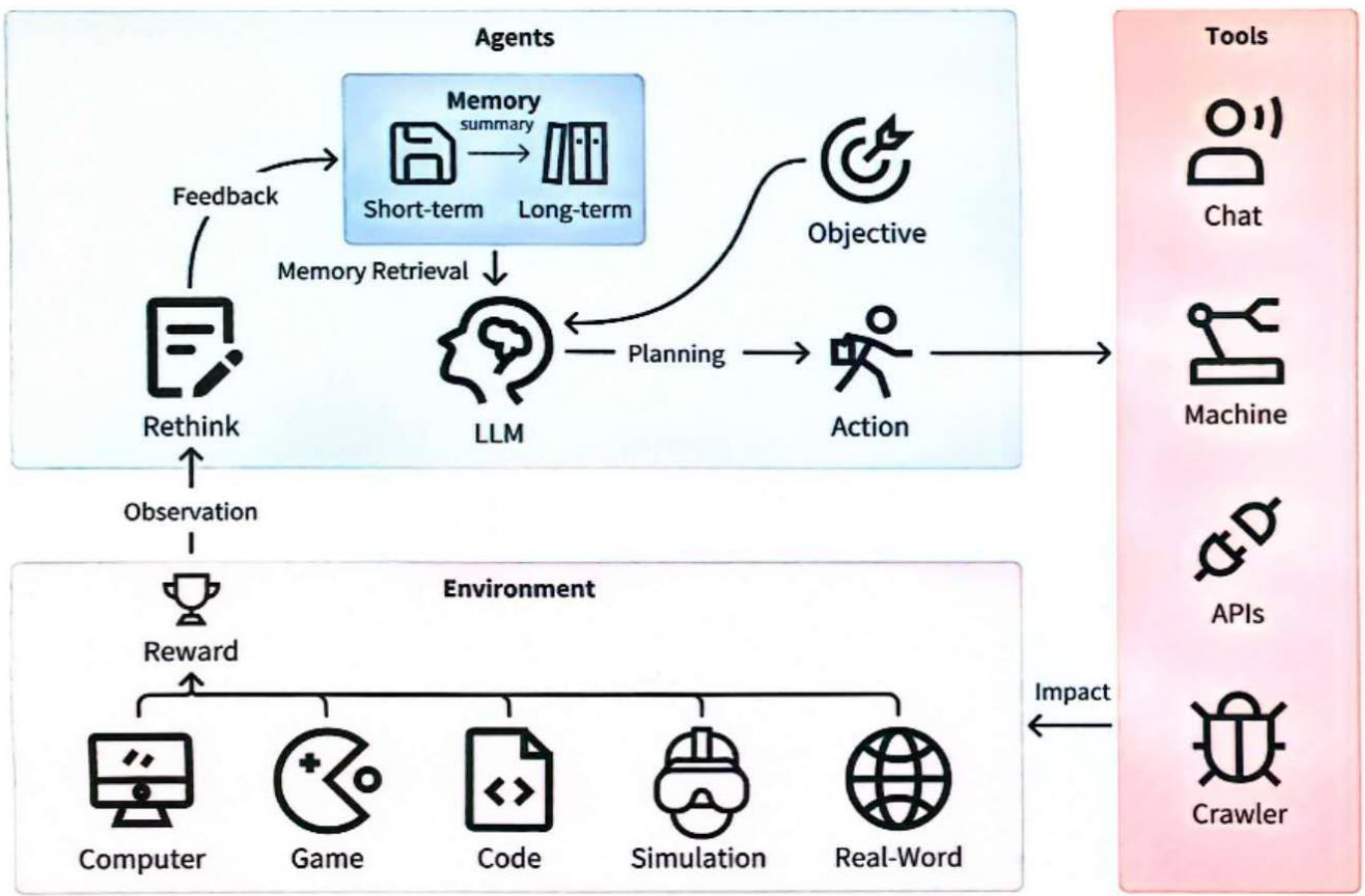
Model-centric AI

- Single inference
- Passive response
- Stateless
- Does not understand goals or objectives
- Does not execute actions

Agent-centric AI

- Long-horizon with multiple tasks
- Proactive planning
- State memorized
- Can understand goals and objectives
- Can execute actions and call (other) tools/agents

Agent: AI4Science



Agent: AI+HEP

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

Agent: AI+HEP

SJTUPA 上海交通大学物理与天文学院

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

What does the Agent do?

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

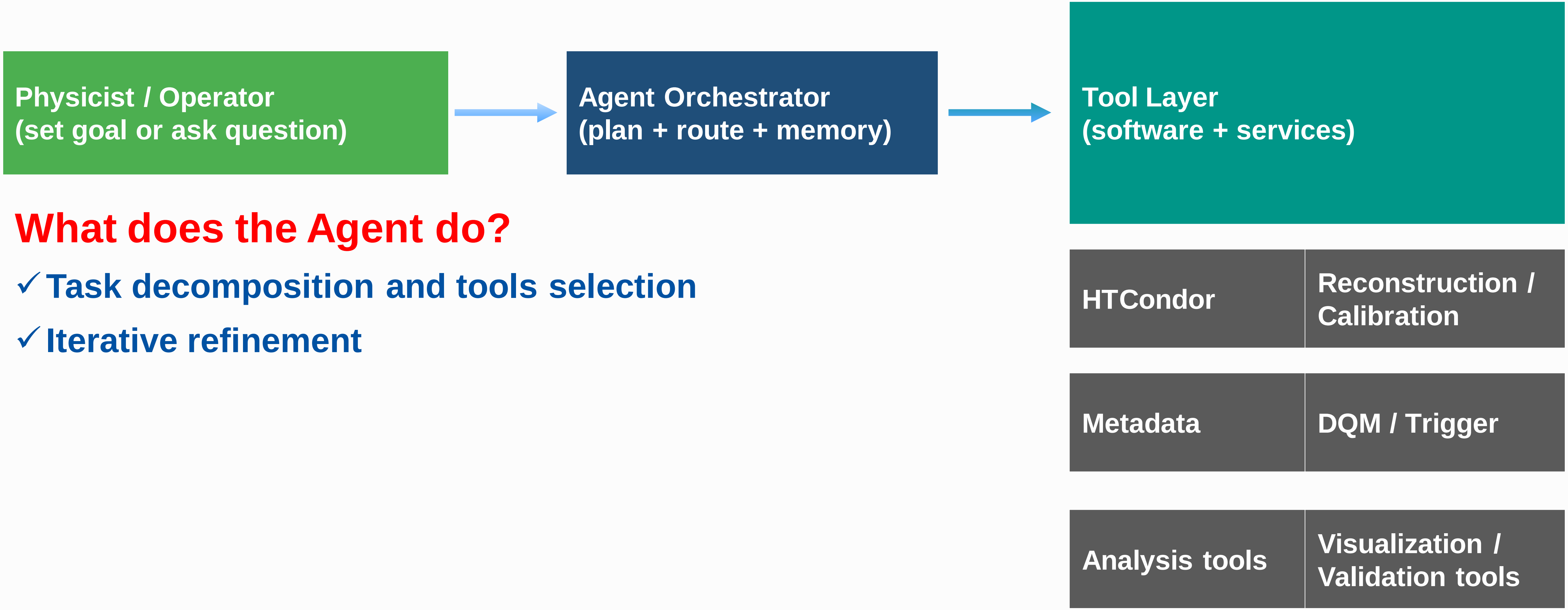
Tool Layer
(software + services)

HTCondor	**Reconstruction / Calibration**
Metadata	**DQM / Trigger**
Analysis tools	**Visualization / Validation tools**

What does the Agent do?

✓ **Task decomposition and tools selection**

Agent: AI+HEP



What does the Agent do?

- ✓ Task decomposition and tools selection
- ✓ Iterative refinement

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

What does the Agent do?

- ✓ Task decomposition and tools selection
- ✓ Iterative refinement
- ✓ Run monitoring: summarize alarms and propose checks

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

What does the Agent do?

- ✓ Task decomposition and tools selection
- ✓ Iterative refinement
- ✓ Run monitoring: summarize alarms and propose checks
- ✓ Launch workflows: submit jobs, monitor DAGs, retry safely

Agent: AI+HEP

SJTUPA 上海交通大学物理与天文学院

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

What does the Agent do?

- ✓ Task decomposition and tools selection
- ✓ Iterative refinement
- ✓ Run monitoring: summarize alarms and propose checks
- ✓ Launch workflows: submit jobs, monitor DAGs, retry safely
- ✓ Data discovery: locate datasets, validate schemas, track provenance

Agent: AI+HEP

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

What does the Agent do?

- ✓ Task decomposition and tools selection
- ✓ Iterative refinement
- ✓ Run monitoring: summarize alarms and propose checks
- ✓ Launch workflows: submit jobs, monitor DAGs, retry safely
- ✓ Data discovery: locate datasets, validate schemas, track provenance
- ✓ Analysis assistance: produce plots and sanity checks with reproducible results

Agent: AI+HEP

SJTUPA 上海交通大学物理与天文学院

Physicist / Operator
(set goal or ask question)

→

Agent Orchestrator
(plan + route + memory)

→

Tool Layer
(software + services)

HTCondor	Reconstruction / Calibration
Metadata	DQM / Trigger
Analysis tools	Visualization / Validation tools

What does the Agent do?

- ✓ Task decomposition and tools selection
- ✓ Iterative refinement
- ✓ Run monitoring: summarize alarms and propose checks
- ✓ Launch workflows: submit jobs, monitor DAGs, retry safely
- ✓ Data discovery: locate datasets, validate schemas, track provenance
- ✓ Analysis assistance: produce plots and sanity checks with reproducible results

Goal: make agents useful by constraining them with tools, permissions, records, validation checks, and reproducible execution.

Conclusion and Outlook

Conclusion and Outlook

✓ **The field of (experimental) high energy physics is rapidly adopting ML and AI techniques leading to real impact on physics results**

Conclusion and Outlook

- ✓ The field of (experimental) high energy physics is rapidly adopting ML and AI techniques leading to real impact on physics results
  - ✓ Diversified use case: classification with both supervised and weakly/self-supervised scheme, reconstruction, simulation and more

Conclusion and Outlook

- ✓ The field of (experimental) high energy physics is rapidly adopting ML and AI techniques leading to real impact on physics results
  - ✓ Diversified use case: classification with both supervised and weakly/self-supervised scheme, reconstruction, simulation and more
  - ✓ Trend: bigger and more sophisticated, more generalized (foundation) model + fine tuning

Conclusion and Outlook

- ✓ The field of (experimental) high energy physics is rapidly adopting ML and AI techniques leading to real impact on physics results
  - ✓ Diversified use case: classification with both supervised and weakly/self-supervised scheme, reconstruction, simulation and more
  - ✓ Trend: bigger and more sophisticated, more generalized (foundation) model + fine tuning
  - ✓ Agent-centric AI is emerging as a potential game-changer

Conclusion and Outlook

- ✓ **The field of (experimental) high energy physics is rapidly adopting ML and AI techniques leading to real impact on physics results**
  - ✓ **Diversified use case: classification with both supervised and weakly/self-supervised scheme, reconstruction, simulation and more**
  - ✓ **Trend: bigger and more sophisticated, more generalized (foundation) model + fine tuning**
  - ✓ **Agent-centric AI is emerging as a potential game-changer**
  - ✓ **Regardless, human insight and expertise remain essential to success in the foreseeable future**

Backup

Self-Supervised: Anomaly Detection

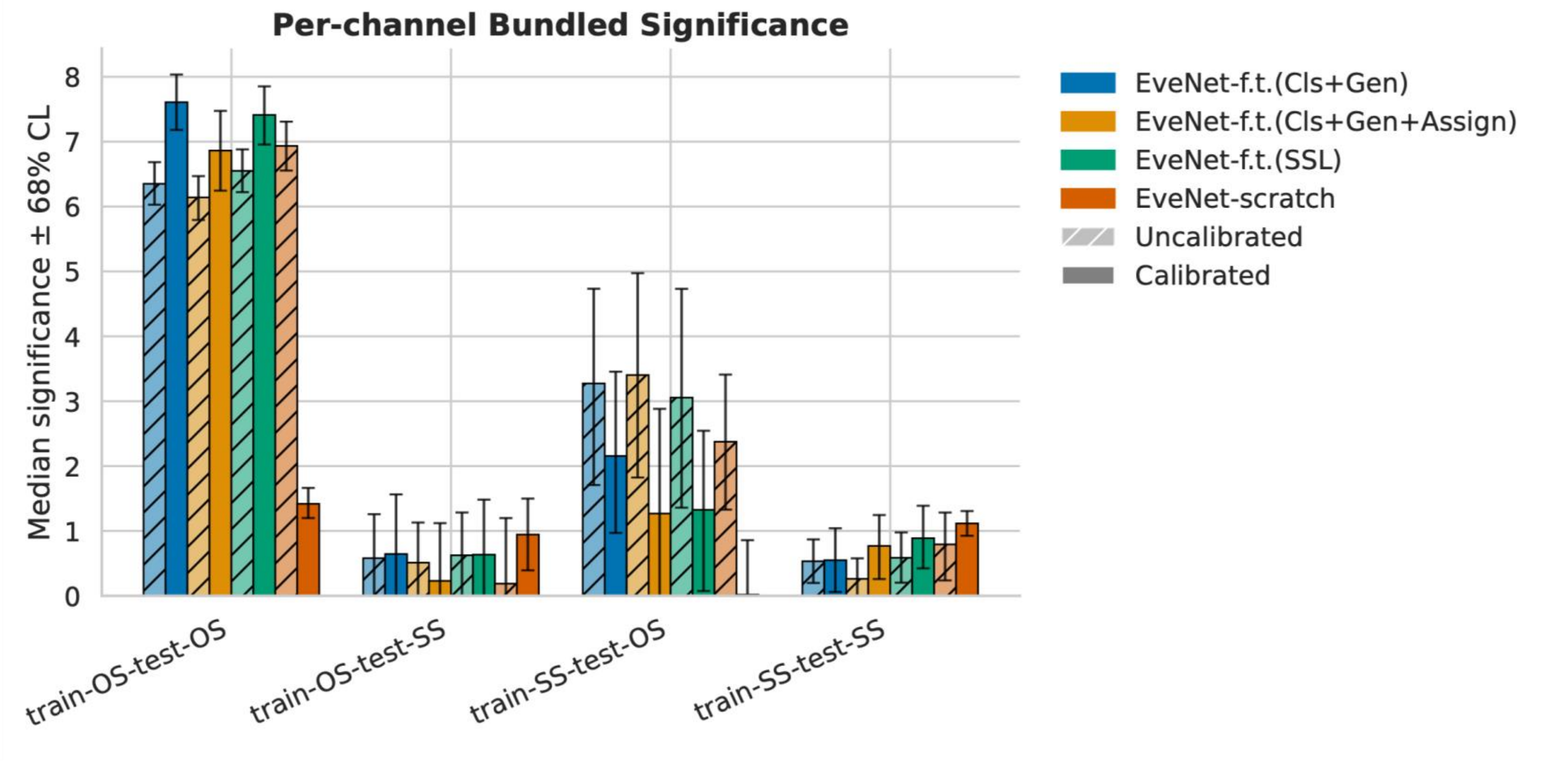
- ***Reference paper***: 2502.14036 (*To test EveNet's generative capability, we extend an existing anomaly detection method **using normalizing flows** by replacing it **with diffusion-based generation** of full 4-momentum)*
- **Dataset**: CMS Open Data (2016 DoubleMu primary dataset) targeting Υ resonances in di-muon final states.

Final Significance (ℓ-reweighting)

- **paper**: $\mathbf{6.4\sigma}$
- **EveNet-Pretrain**: $\mathbf{7.5\sigma}$
- ~~**EveNet-Scratch**: $\mathbf{?\sigma}$~~ (**mass sculpting** ✗)

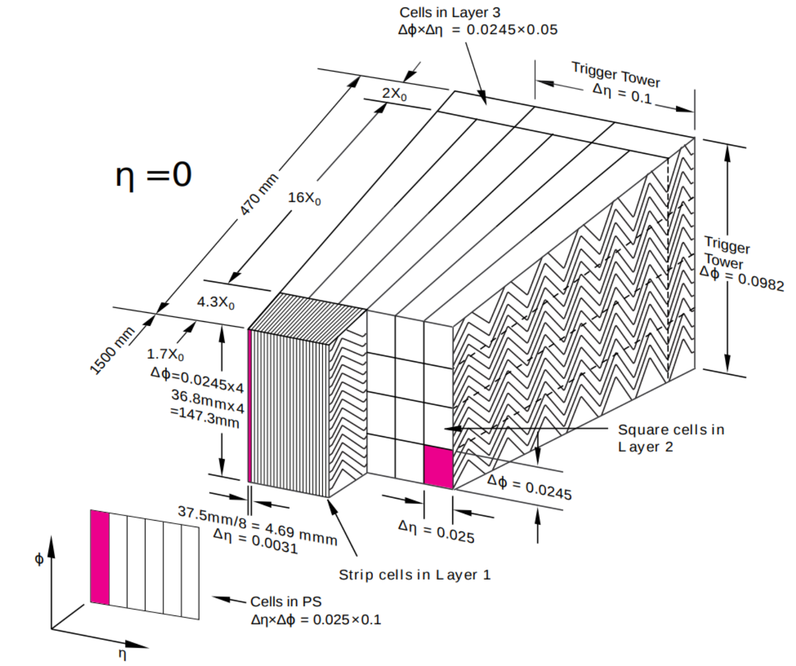
Note: *the energy regime here is even different from the main samples in pretrain*

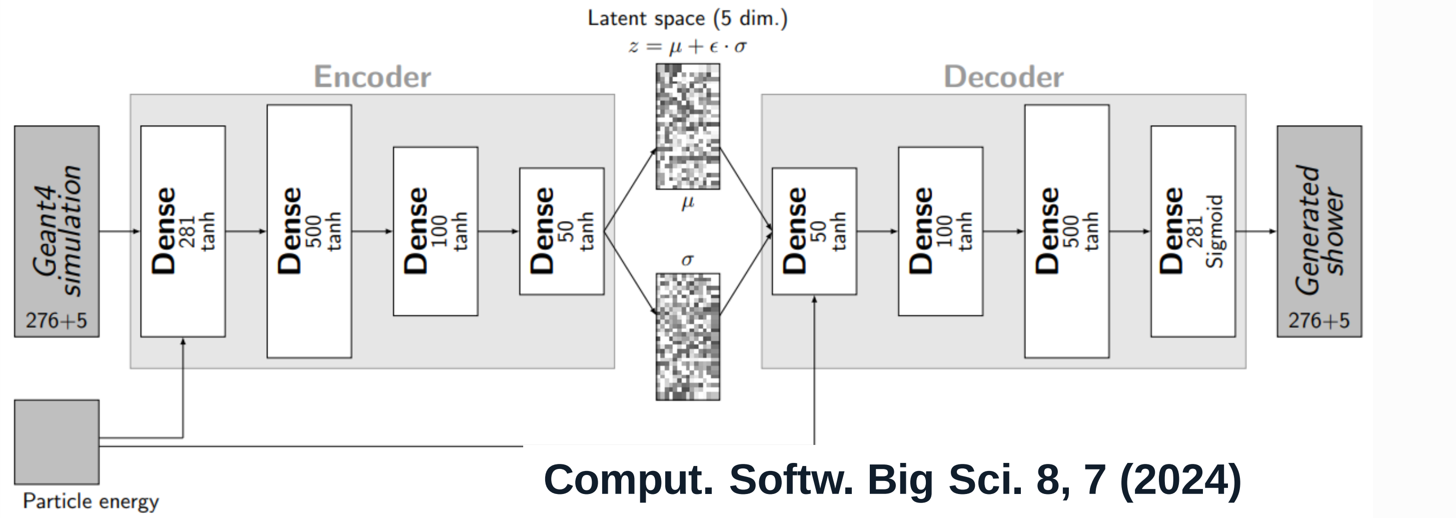
$\mathbf{1\sigma}$ **improvement** on significance



✓ **Trained on Self-Supervised Generation task**

Faster Simulation: Photon Showers





Comput. Softw. Big Sci. 8, 7 (2024)

Use Variational autoencoders (VAE) and generative adversarial networks (GAN) to simulation ATLAS photon showers

- **VAE/GAN: x100 faster than GEANT4 full simulation**
- **Good agreement between GAN/VAE and Geant4 for EM showers of different energies**
- **GAN needs improvement in the longitudinal shower development**