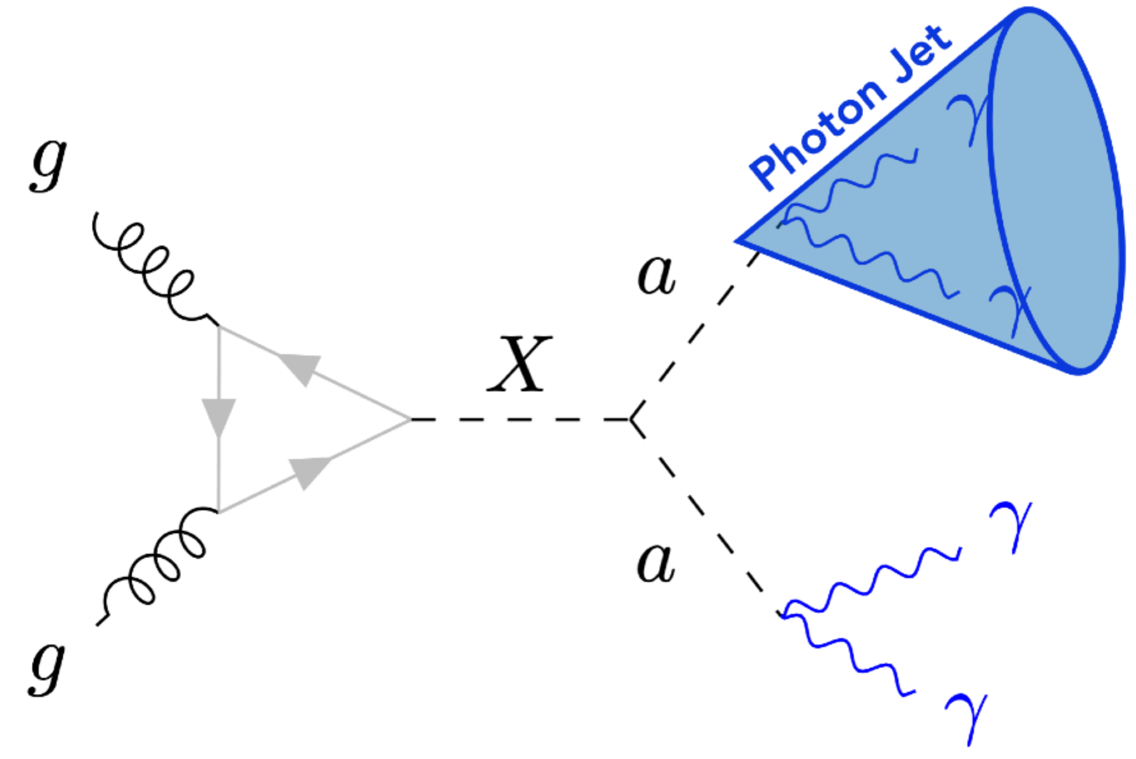
Transformer for Collimated Photon Identification Using Low-Level Calorimeter Signals

Gabriel Matos, o.b.o. authors
Columbia University
September 15th, 2026

ML4Jets 2026, Vienna, Austria

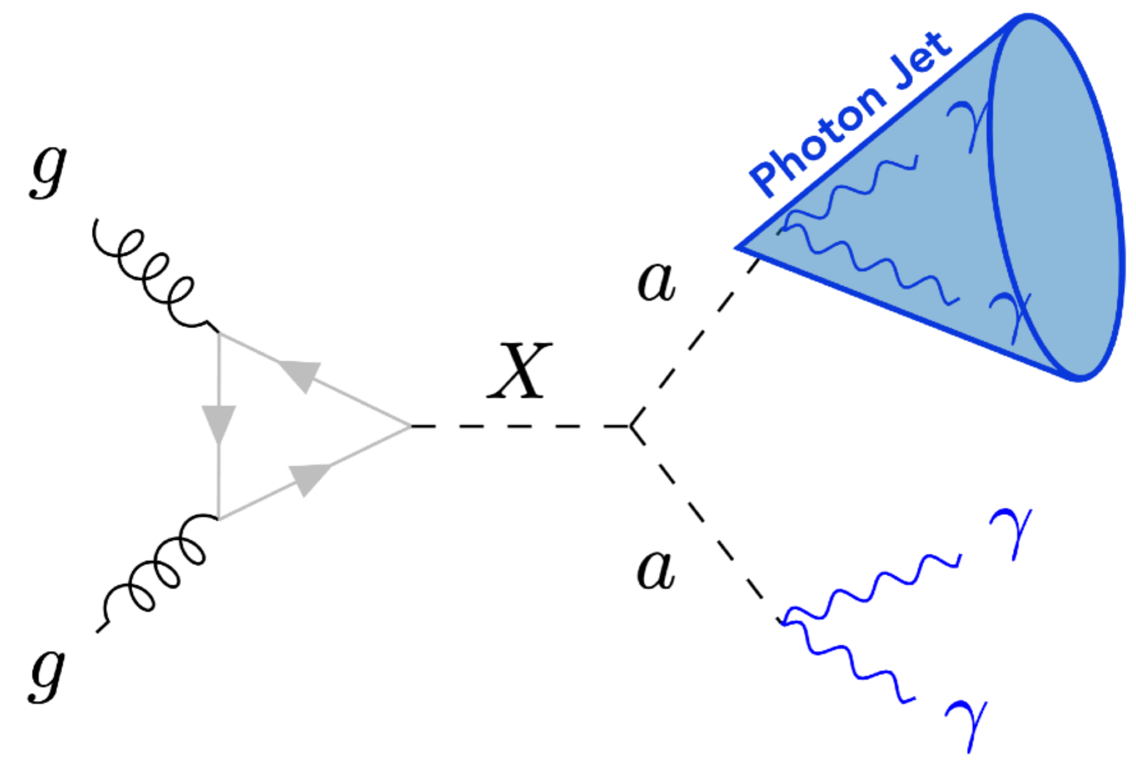
Motivation for Overlapping Photons

- Various extensions to the SM predict the existence of light (pseudo)scalar, **axion-like particles (ALPs)** with a large branching ratio to photons
  - At colliders, ALPs can be produced from a Higgs boson or new heavy scalar X, where for sufficiently low ALP mass, the $a \to \gamma\gamma$ decay is boosted



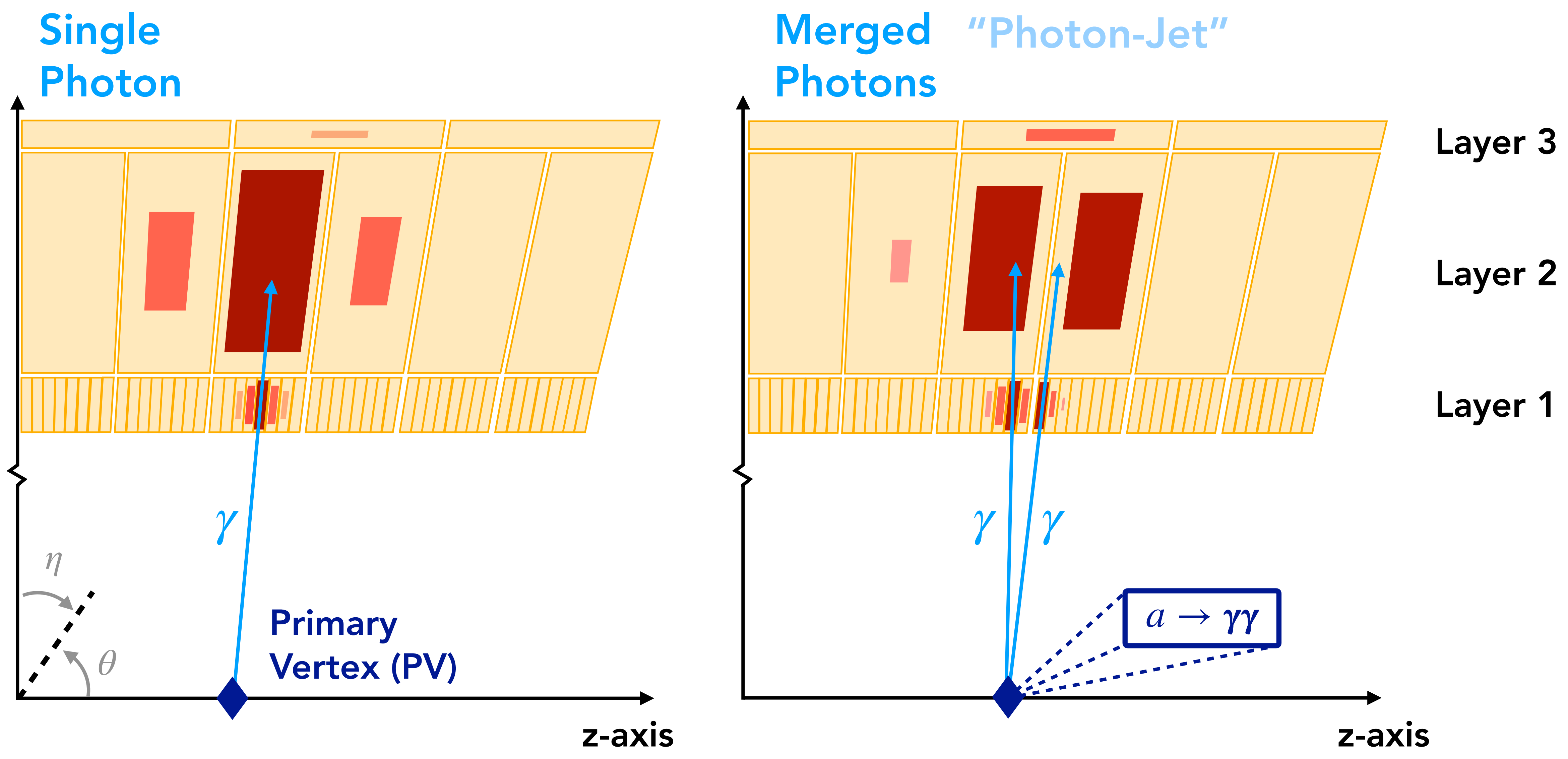
Motivation for Overlapping Photons

- Various extensions to the SM predict the existence of light (pseudo)scalar, **axion-like particles (ALPs)** with a large branching ratio to photons
  - At colliders, ALPs can be produced from a Higgs boson or new heavy scalar X, where for sufficiently low ALP mass, the $a \to \gamma\gamma$ decay is boosted



$$\Delta R_{\gamma\gamma} \equiv \sqrt{\left(\Delta\eta_{\gamma\gamma}\right)^2 + \left(\Delta\phi_{\gamma\gamma}\right)^2} \simeq \frac{2m_a}{p_{\mathrm{T},a}} \simeq \frac{4m_a}{m_X}$$

- For collimations of $\Delta R_{\gamma\gamma} \lesssim \mathcal{O}(10^{-2})$, the γγ pair is **reconstructed in ATLAS as a single photon**
- We call these **photon jets (PJs)**: overlapping photons that mimmic isolated photons

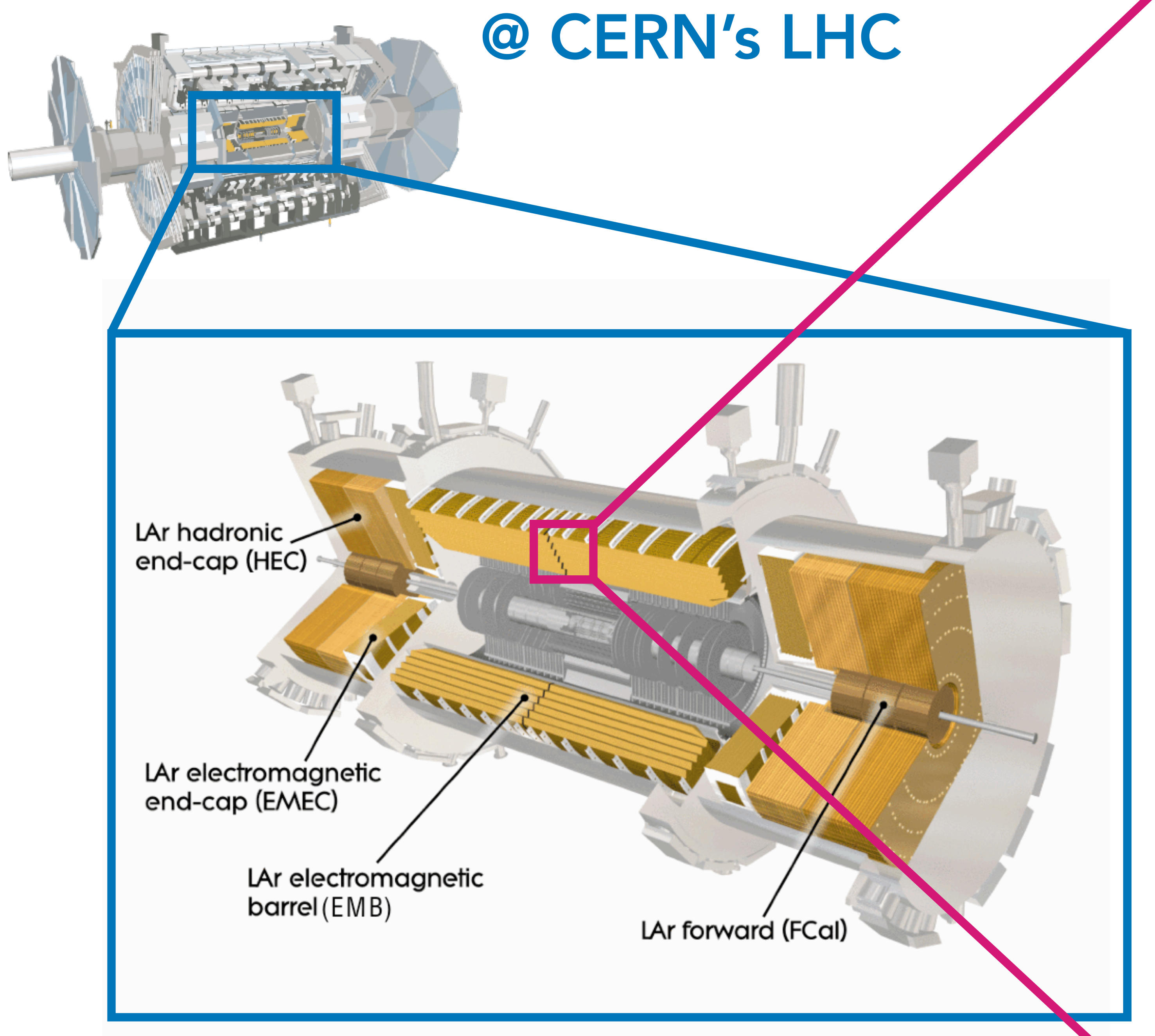


ATLAS liquid argon calorimeter

ATLAS Liquid Argon Calorimeter

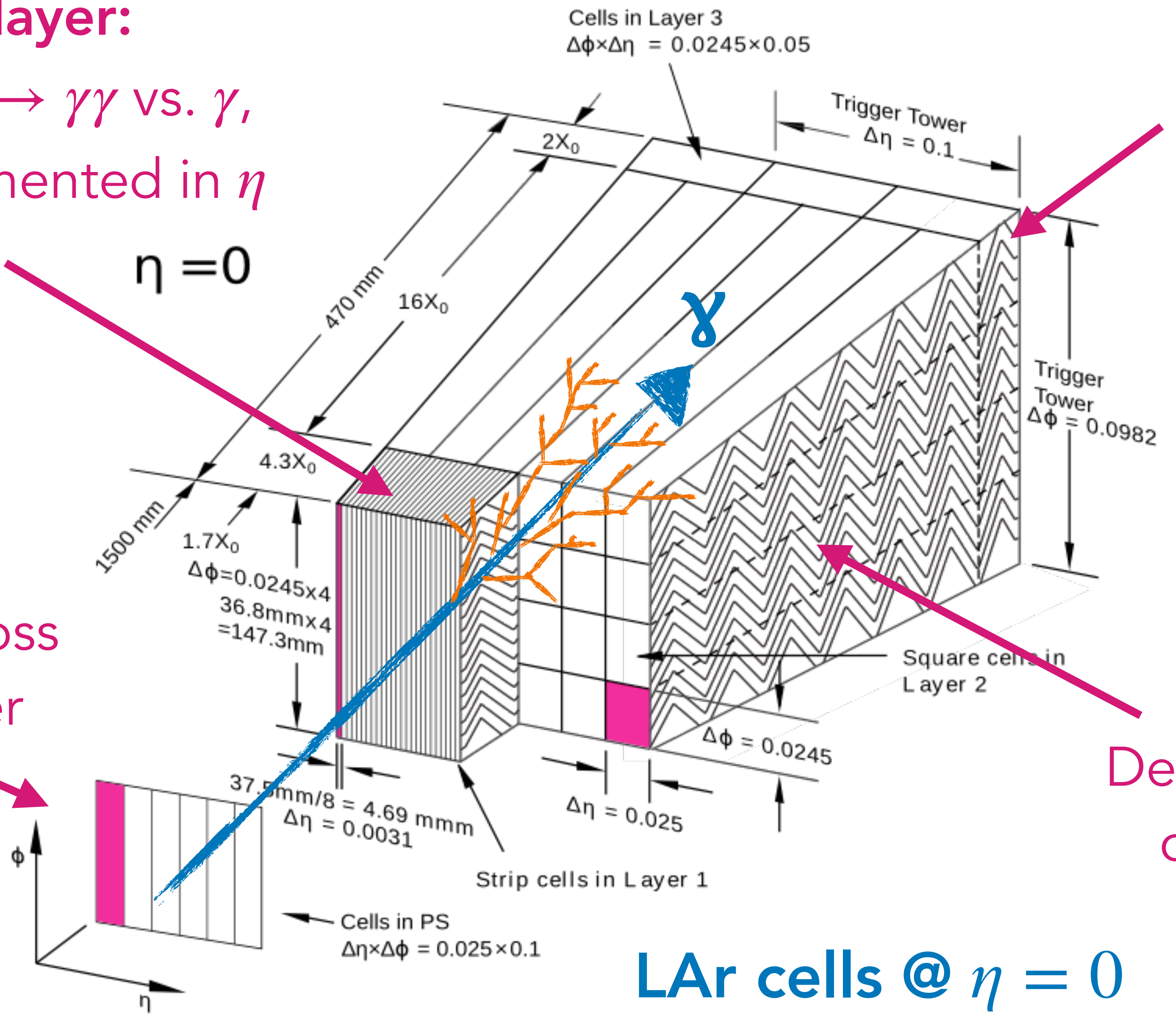
- In ATLAS, photons are detected by the liquid argon (LAr) electromagnetic calorimeters

ATLAS Experiment @ CERN's LHC



(1) Front layer: Distinguish $\pi^0 \to \gamma\gamma$ vs. γ, very finely segmented in η

(0) Presampler: Measures energy loss before calorimeter



(3) Back layer: Catches the tail of the shower

(2) Middle layer: Deep, collects the majority of the shower's energy

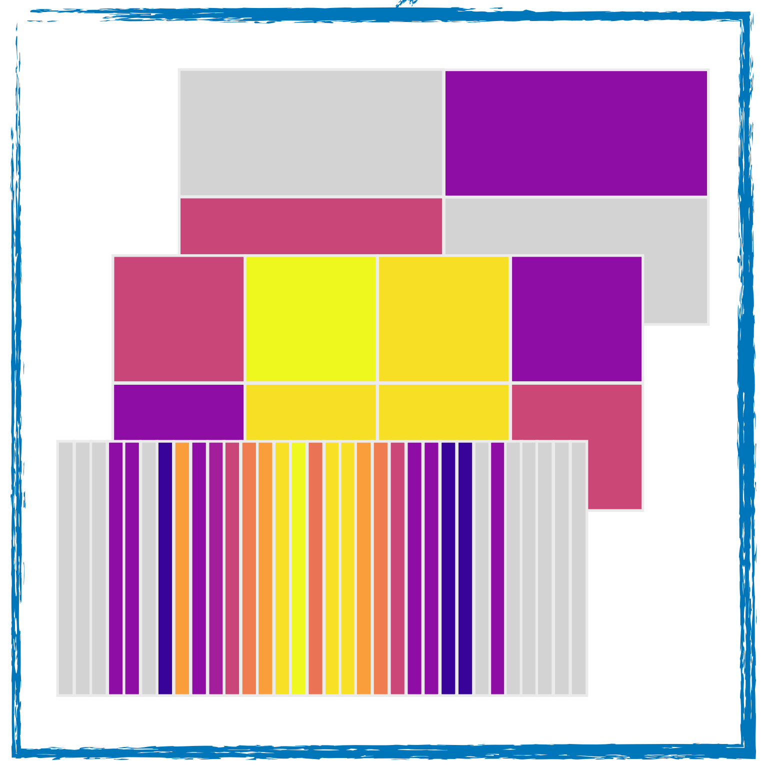
LAr cells @ $\eta = 0$

- The calorimeter is segmented by ~185k **cells**, with a finest segmentation of $\Delta\eta \sim 0.0031$ in the first layer strips

*CMS has published approaches with CNNs [1, 2] and GNNs [3]
[1] Phys. Rev. D 108 (2023) 052002 [link]
[2] Phys. Rev. Lett. 134 (2025) 041801 [link]
[3] JHEP 07 (2026) 119 [link]

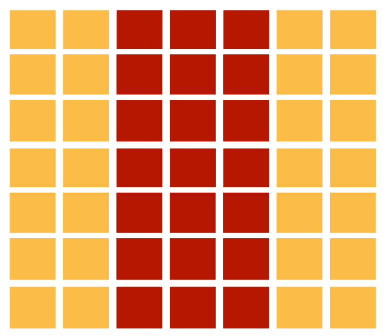
Classifying Photon Jets

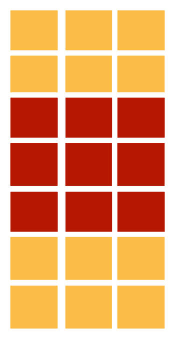
- Previous discriminants for PJs vs. single photons in ATLAS built on **shower shape variables (SSVs)***



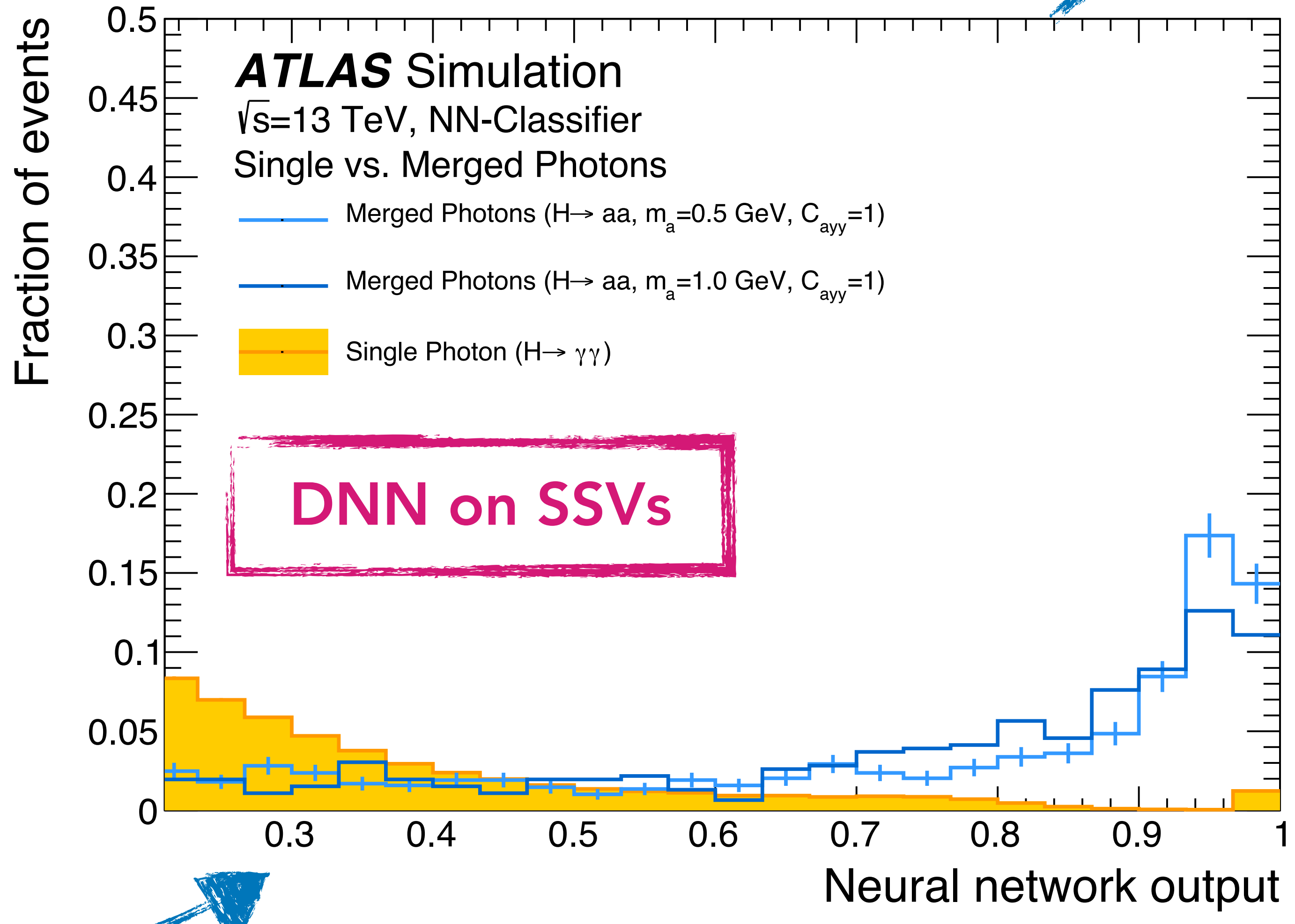
Energy deposits in cells

$$R_\eta = \frac{E^{L2}_{3\times7}}{E^{L2}_{7\times7}} \qquad R_\phi = \frac{E^{L2}_{3\times3}}{E^{L2}_{3\times7}}$$

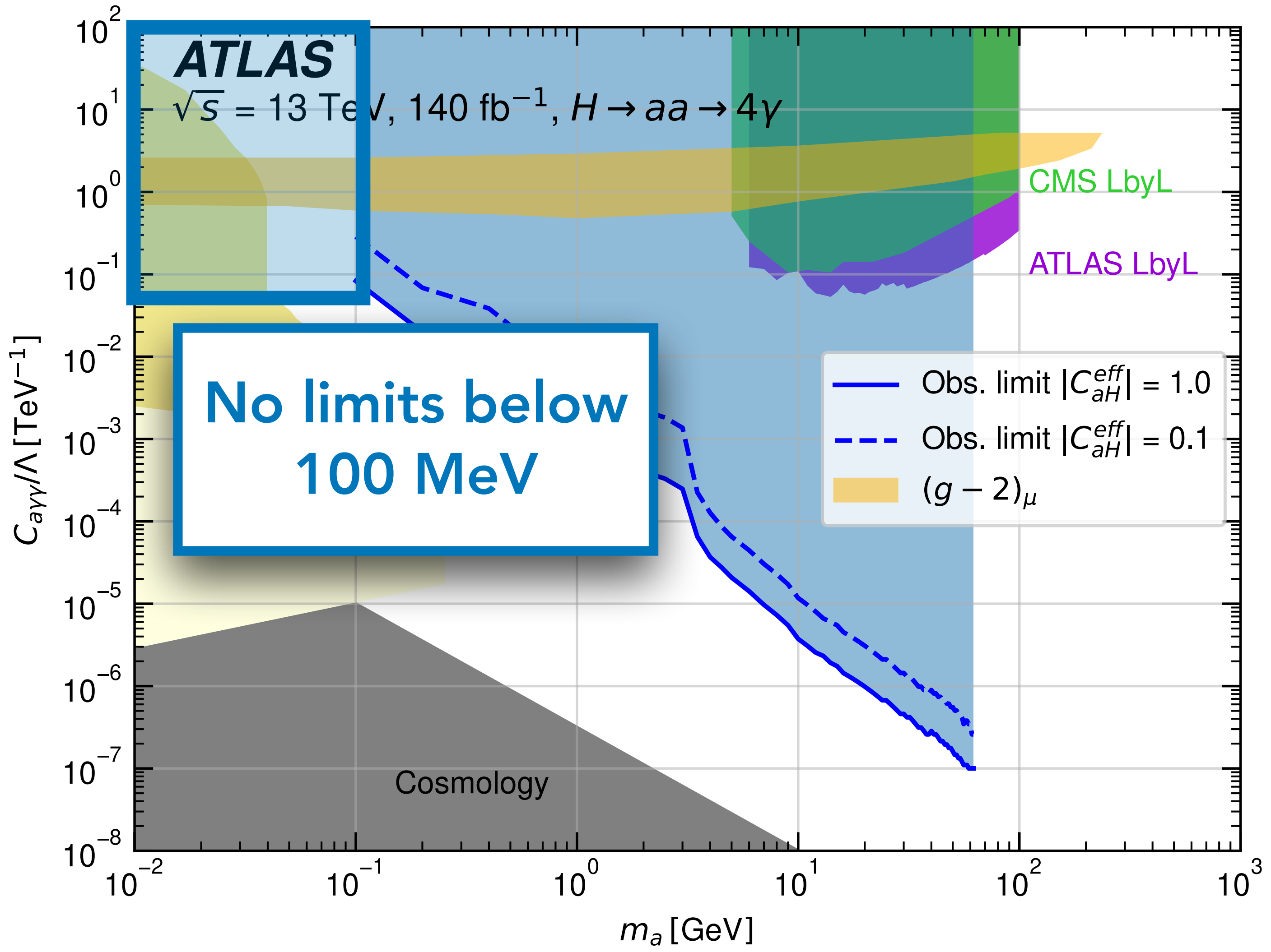




$$f_{\text{side}} = \frac{E^{L1}_{7} - E^{L1}_{3}}{E^{L1}_{3}}$$



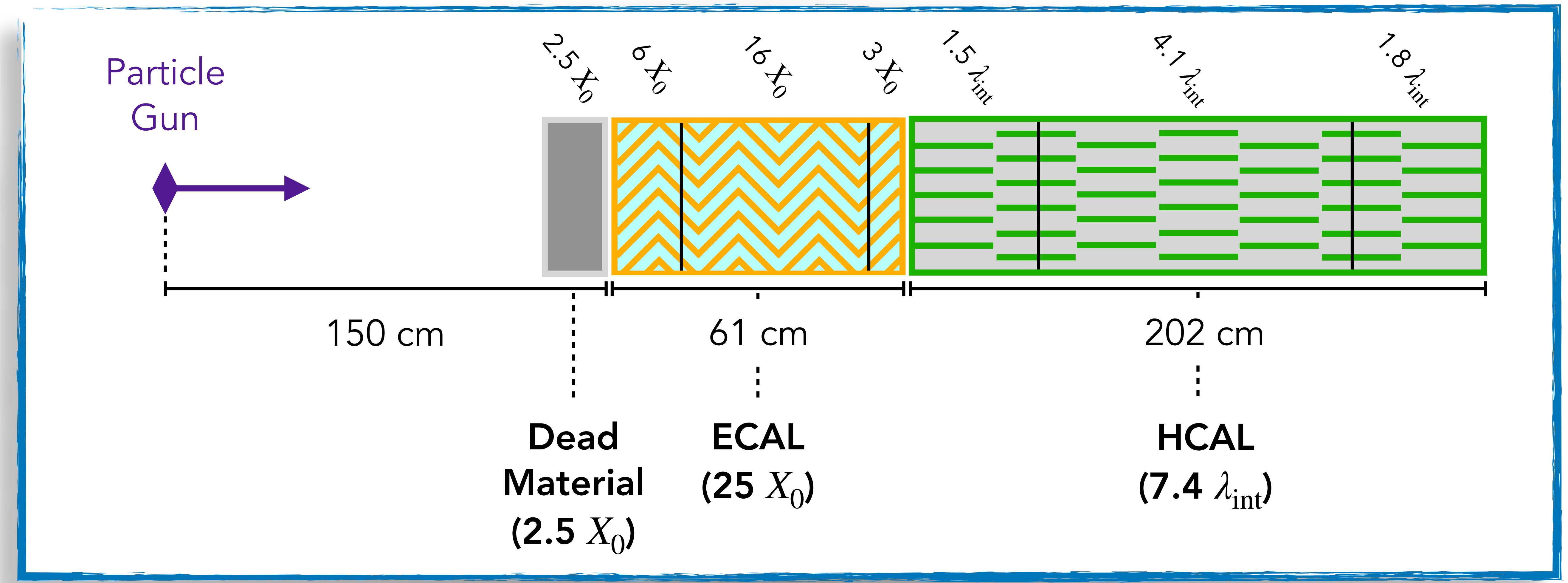
ATLAS $H \to aa \to 4\gamma$ Run 2 Analysis [link]



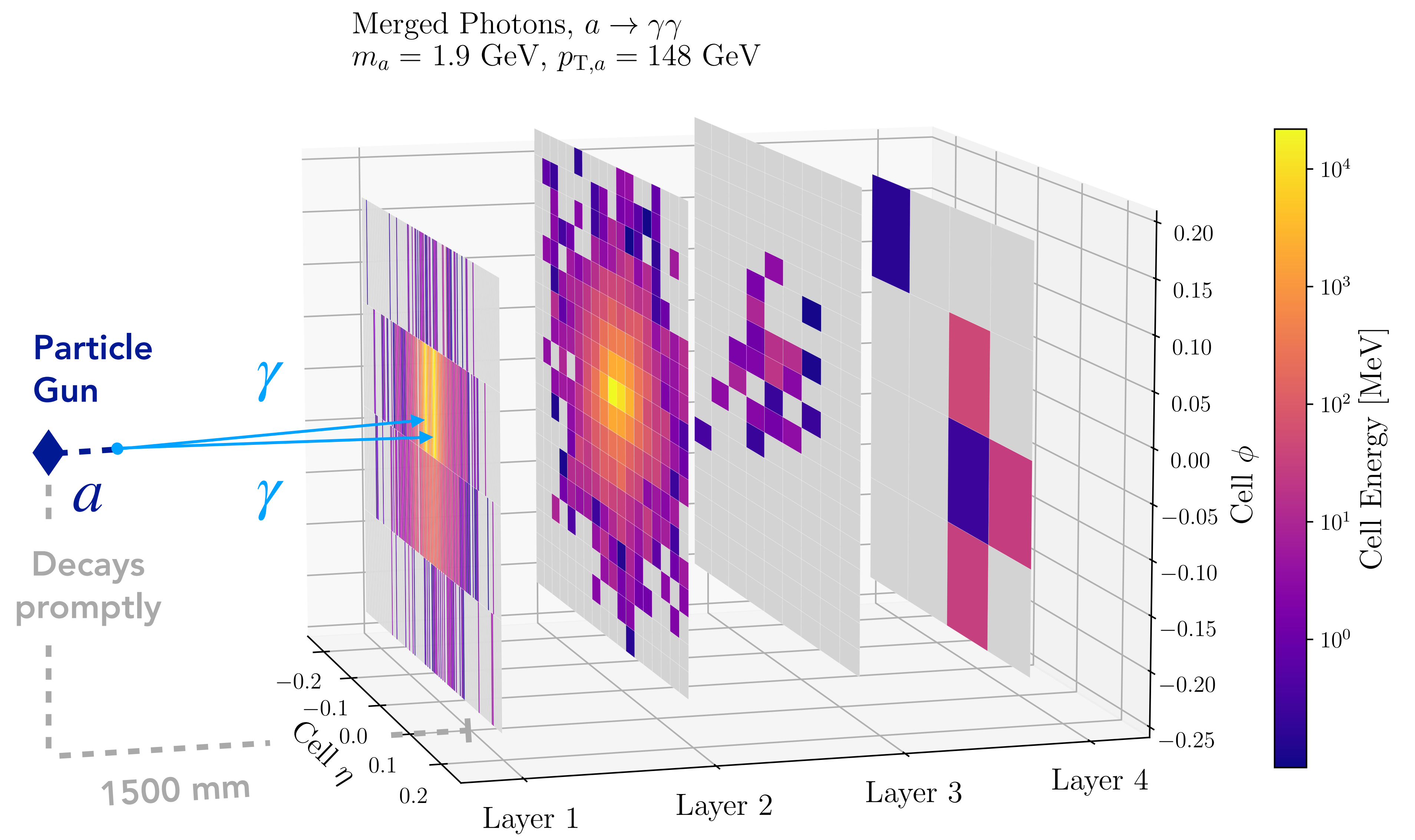
- **SSVs provide only a coarse resolution of the shower development for collimated topologies**
- **Low-level modeling** with ML directly with the **calorimeter cells** can provide x10 more granularity
  - Modeling with the cells can also provide additional handles like the invariant mass of the photon pair, which can give discriminating power to help control the background from fakes (e.g. $\pi^0/\eta \to \gamma\gamma$)

ATLAS-like Calorimeter for ML Development

- To assess the performance gains of using the **cells**, we designed an exploratory study using a simplified, **ATLAS-like calorimeter** for fast simulation of the cell-level data
  - Benchmark dataset produced with GEANT4 using a **particle gun** shooting single photons and photon jets (from $a \to \gamma\gamma$), used for ML development of a cell-level, PJ classifier



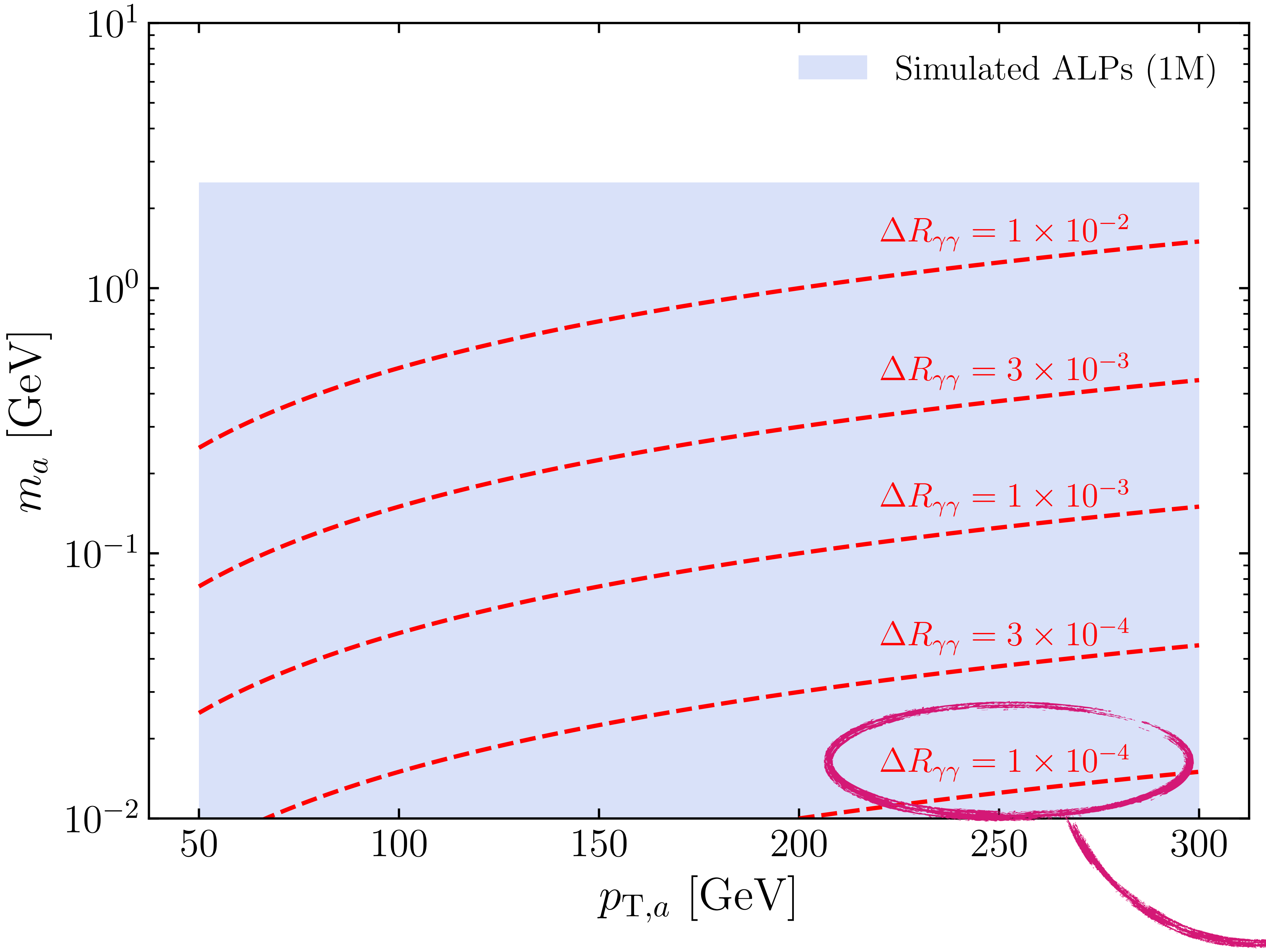
Matches granularity of ATLAS at $\eta = 0$ (without pre-sampler)



Simulation provides cell-level response in ECAL + HCAL

Benchmark Dataset

- Generated ALPs uniformly cover masses $m_a \in (0.01, 2.5)$ GeV and momentum $p_T \in (50, 300)$ GeV to provide varying levels of collimation between the photon pairs
  - Single photons generated to uniformly cover the same momentum range



Dataset	Train	Validation	Test	Total
ALPs	800k	100k	100k	1M
Single Photons	800k	100k	100k	1M
Total	**1.6M**	**200k**	**200k**	**2M**

Train/test/validation split of 80/10/10

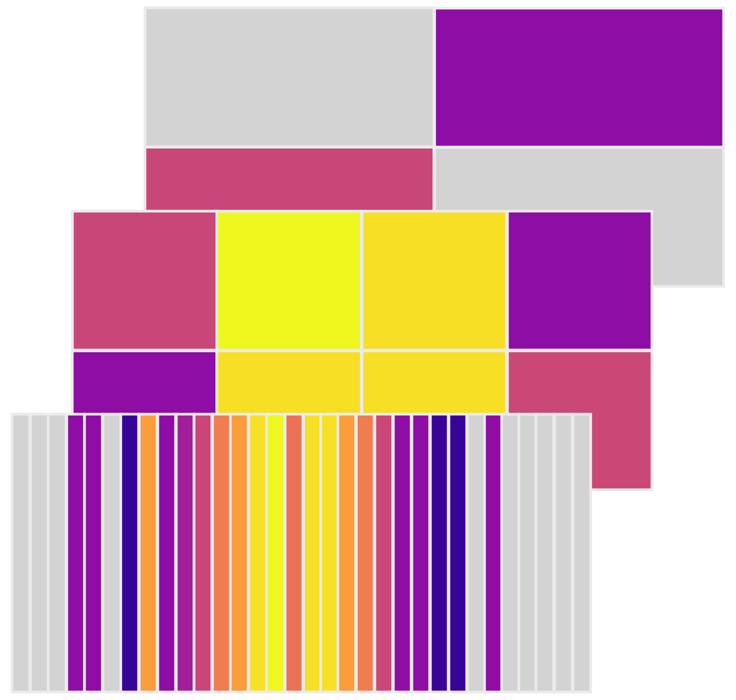
ALP masses down to **10 MeV** and $\mathbf{\Delta R_{\gamma\gamma} \sim 10^{-4}}$ separation to cover range not previously targeted by ATLAS

Transformers for Cell-Level Learning

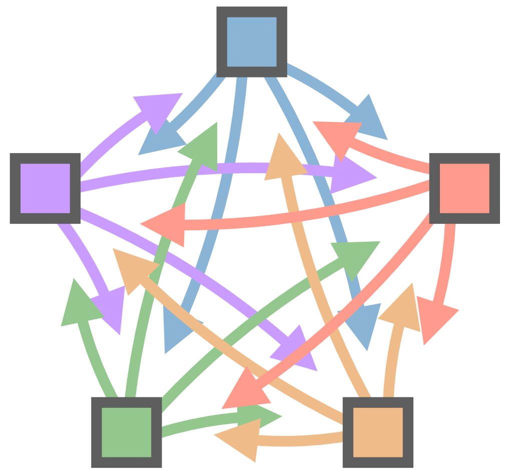
- Cell-level ML requires **variable-length, set-like inputs** due to variable detector granularity and number of cells, while modeling **pairwise correlations** between cells from the collimated shower

Architecture	Variable-length, set-modeling	Explicit pairwise interactions
BDT/DNN	✗	✗
CNN	✗	✓
DeepSets	✓	✗
Transformer*	✓	✓

*Transformers show promising performance for low-level modeling in ATLAS e.g. ATLAS flavor tagging, e/ɣ energy calibration



Cell tokens mapped to keys (K), queries (Q), and values (V)



Each cell attends to every other cell via multi-headed self-attention (MHSA)

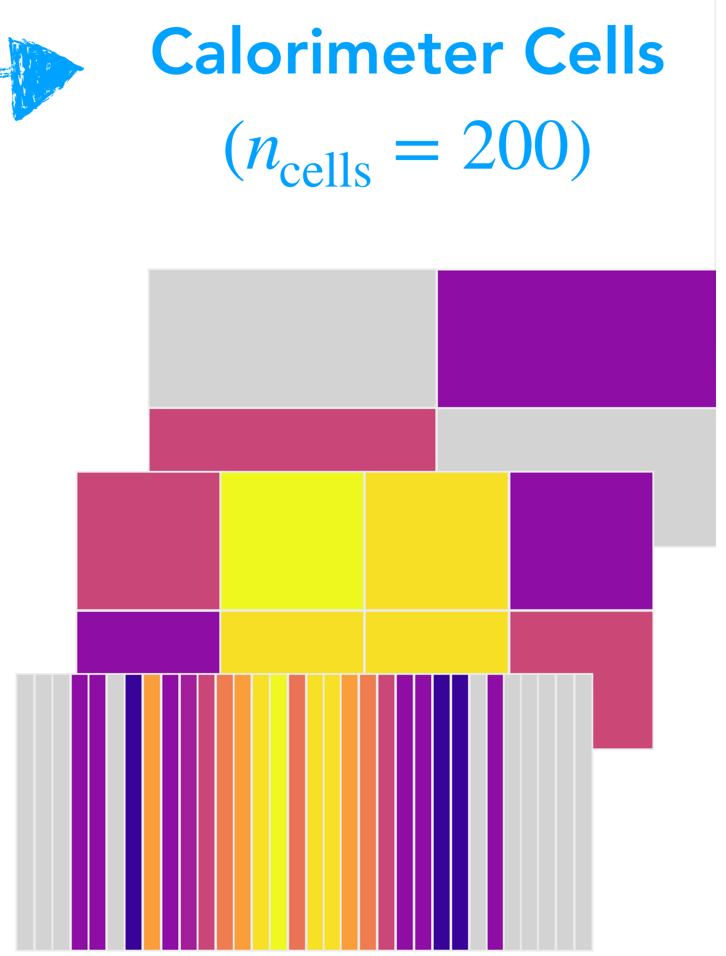
$$\mathrm{MHSA} = \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V \times \mathrm{Heads}$$

Explicit modeling of pairwise correlations between cells in the calorimeter

Transformer for Photon Jet Classification

Input cell energies and positions

n_{cells} chosen to match cell multiplicity in ATLAS topoclusters

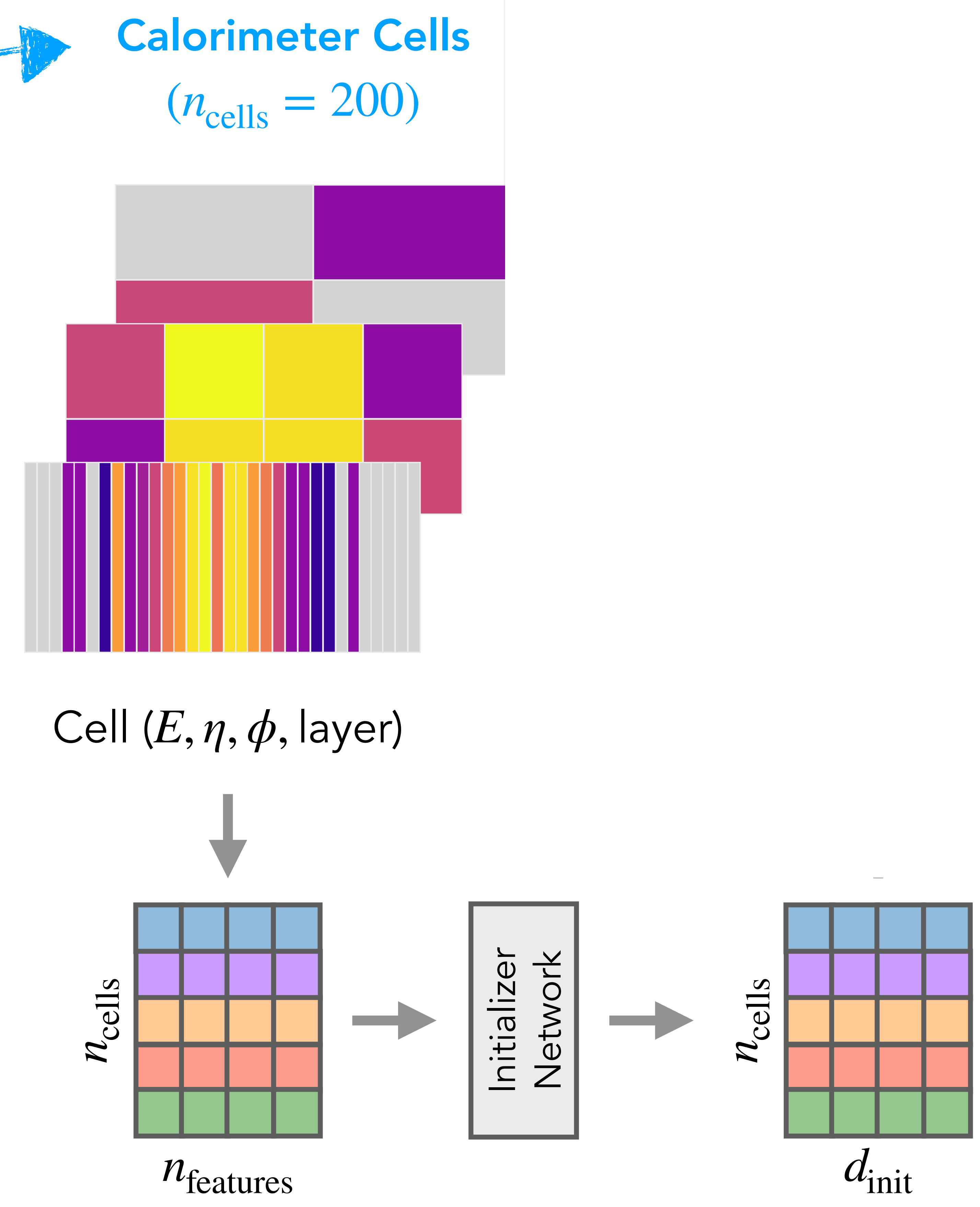


Cell $(E, \eta, \phi,$ layer)

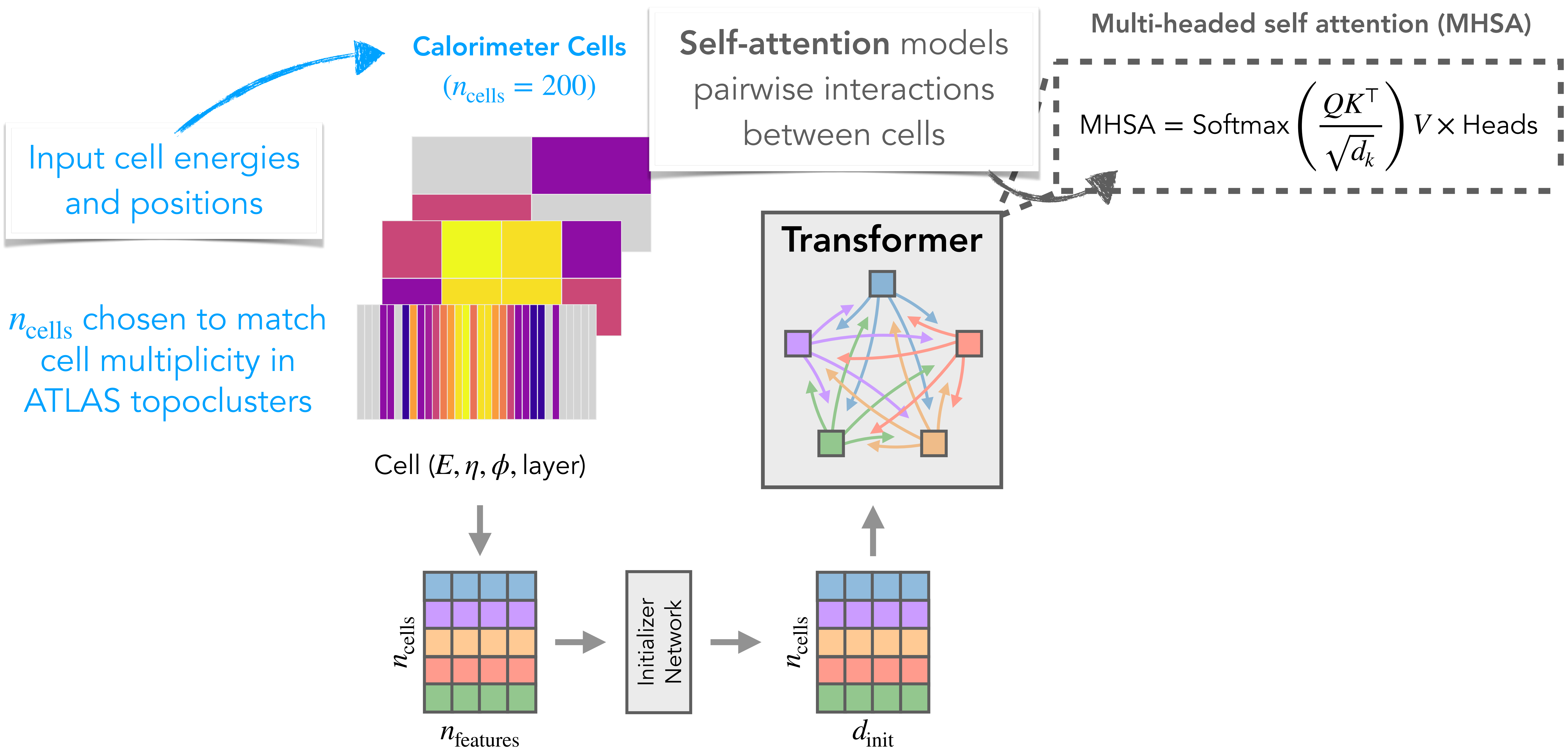
Transformer for Photon Jet Classification

Input cell energies and positions

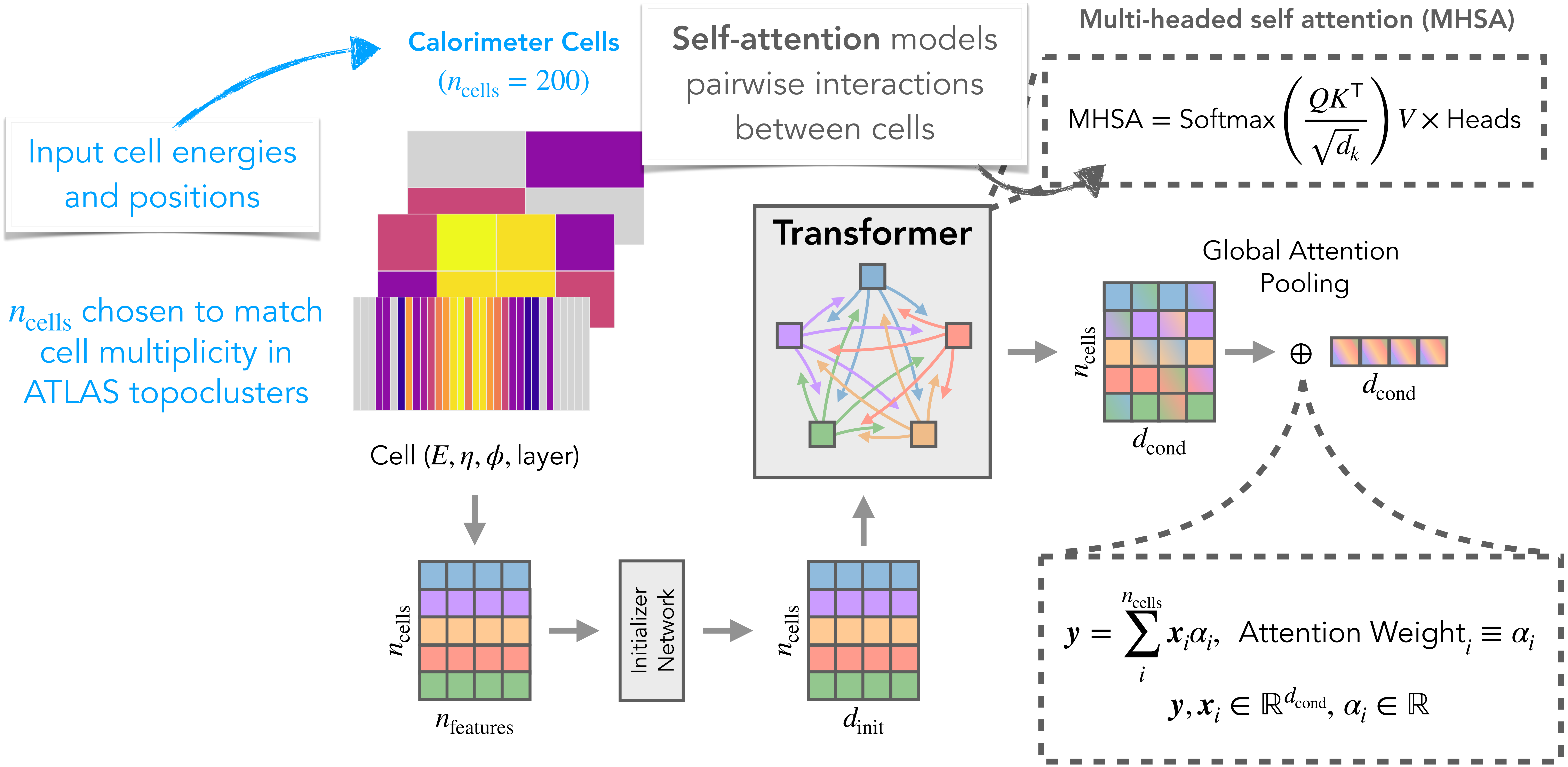
n_{cells} chosen to match cell multiplicity in ATLAS topoclusters



Transformer for Photon Jet Classification



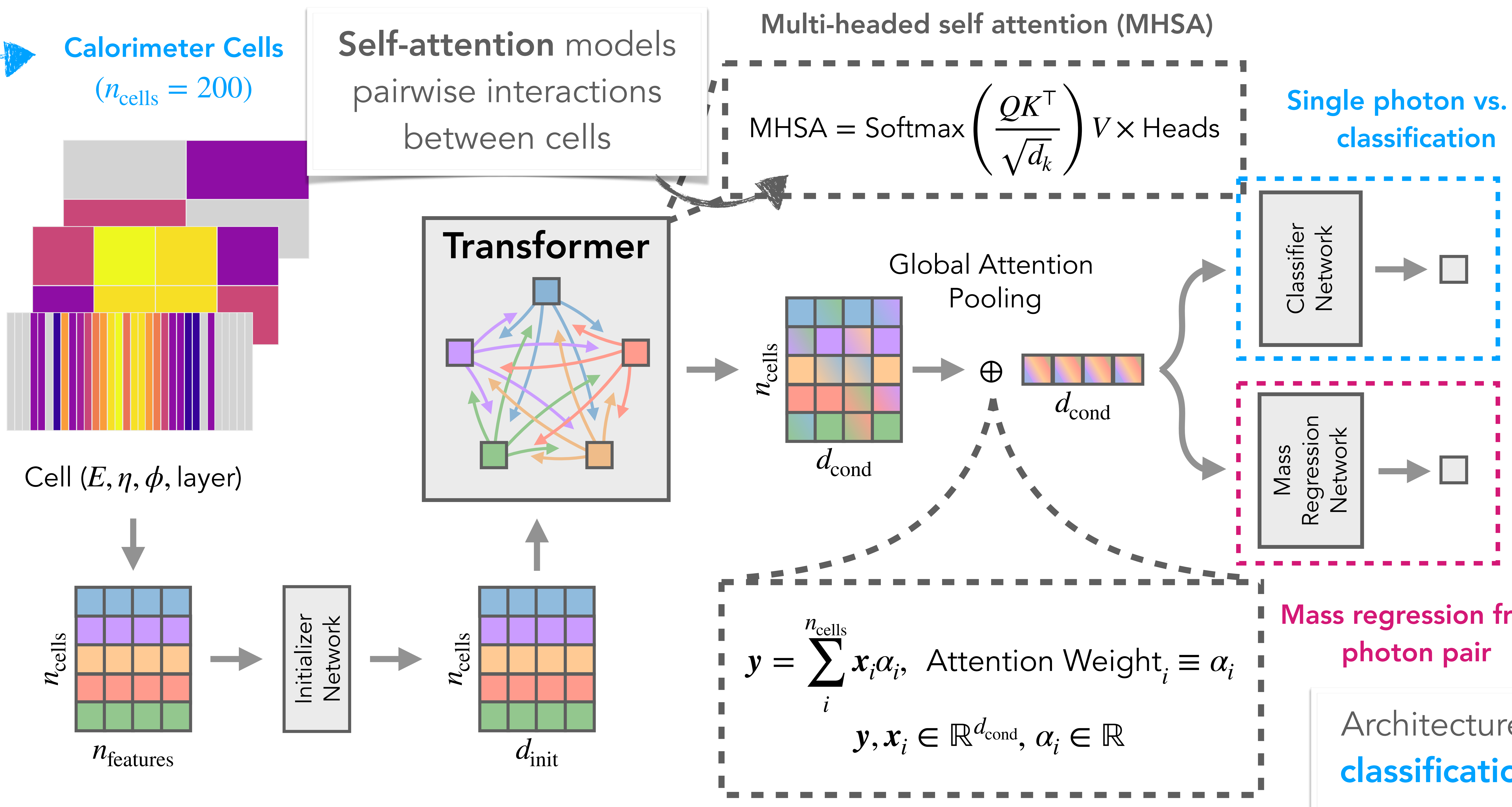
Transformer for Photon Jet Classification



Transformer for Photon Jet Classification

Input cell energies and positions

n_{cells} chosen to match cell multiplicity in ATLAS topoclusters



Architecture performs both **classification** and **invariant mass regression**

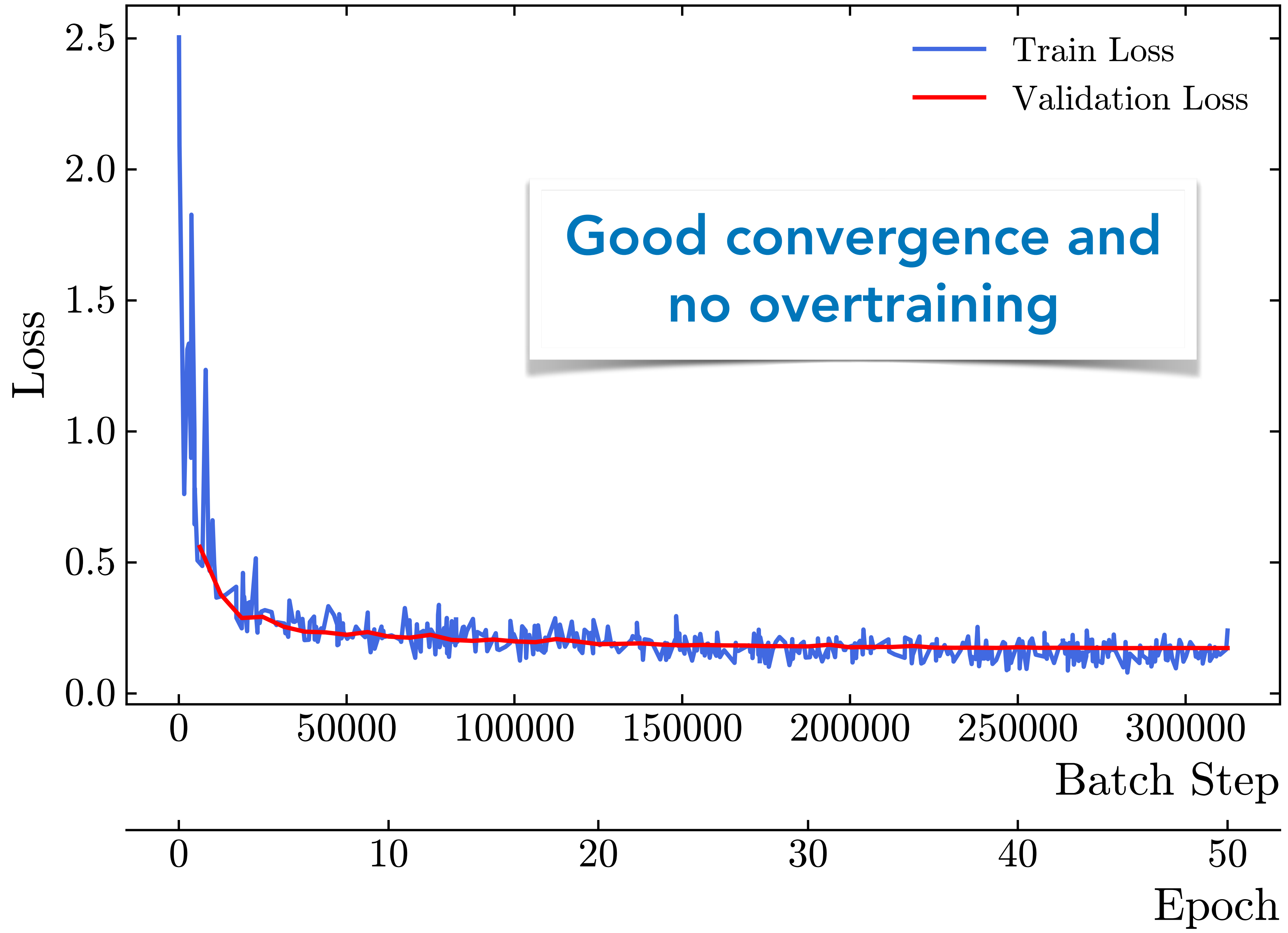
Transformer Parameters & Training

- Transformer implemented and trained using ATLAS's flavor tagging SALT [link] framework

$$\mathscr{L} = \underbrace{\mathrm{BCE}}_{\text{Classification}} + \lambda \sum_i \underbrace{\log\cosh(y_{\mathrm{pred},i} - y_i)}_{\text{Mass regression}}$$

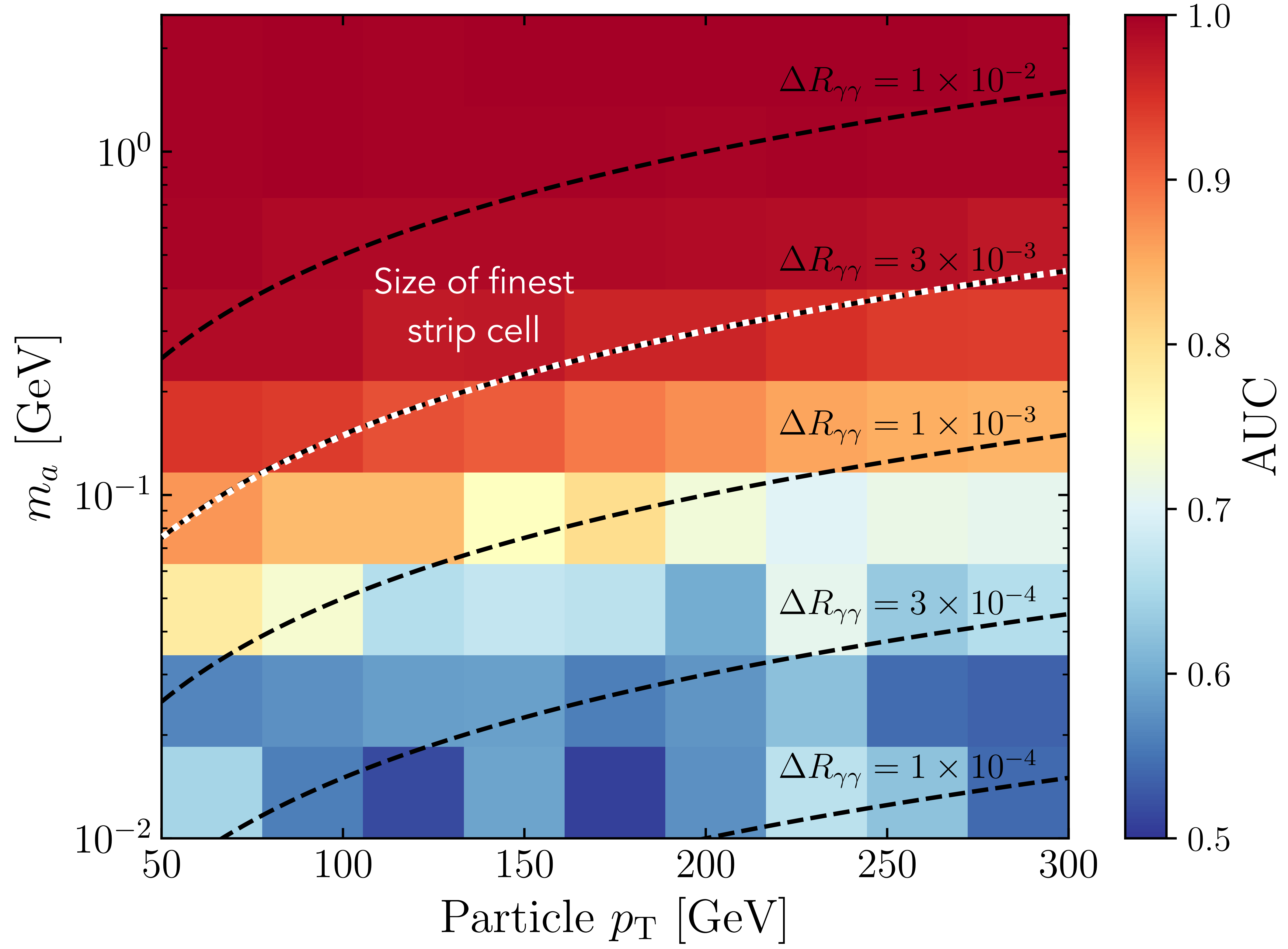
- Log cosh chosen to provide stability due to regression over ~2 orders of magnitude
- $\lambda = 4$ is chosen to roughly give equal weight to classification and mass regression in the loss

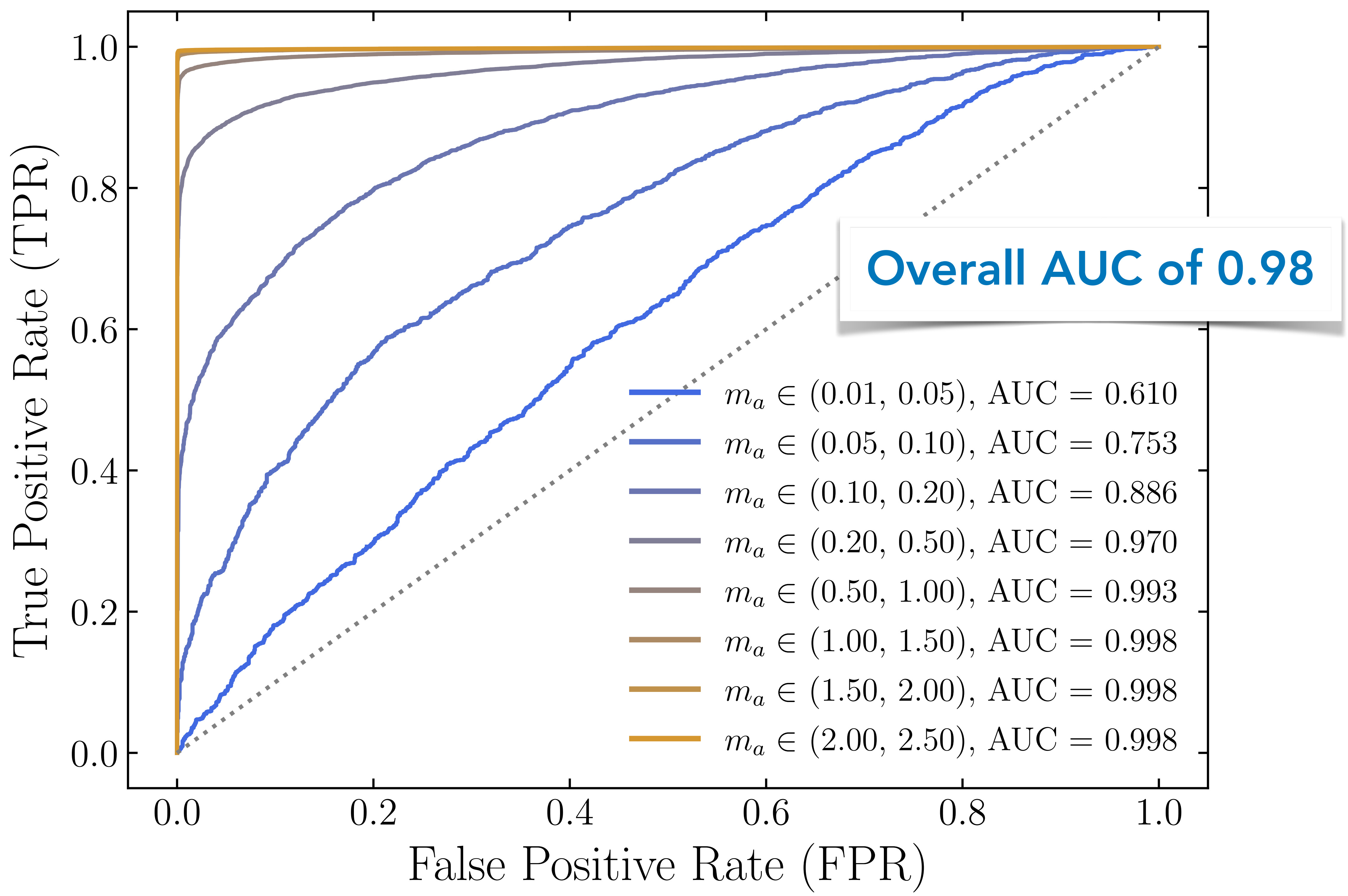
Component	Parameters
Input size	Cell (E, η ϕ, layer): 4 x 200
Input encoder	[256, 128]
Transformer encoder	8 attention heads, [128] x 4
Classification	[128, 64, 32]
Mass regression	[128, 64, 32]
Number of parameters	**634k**
Epochs	**50**



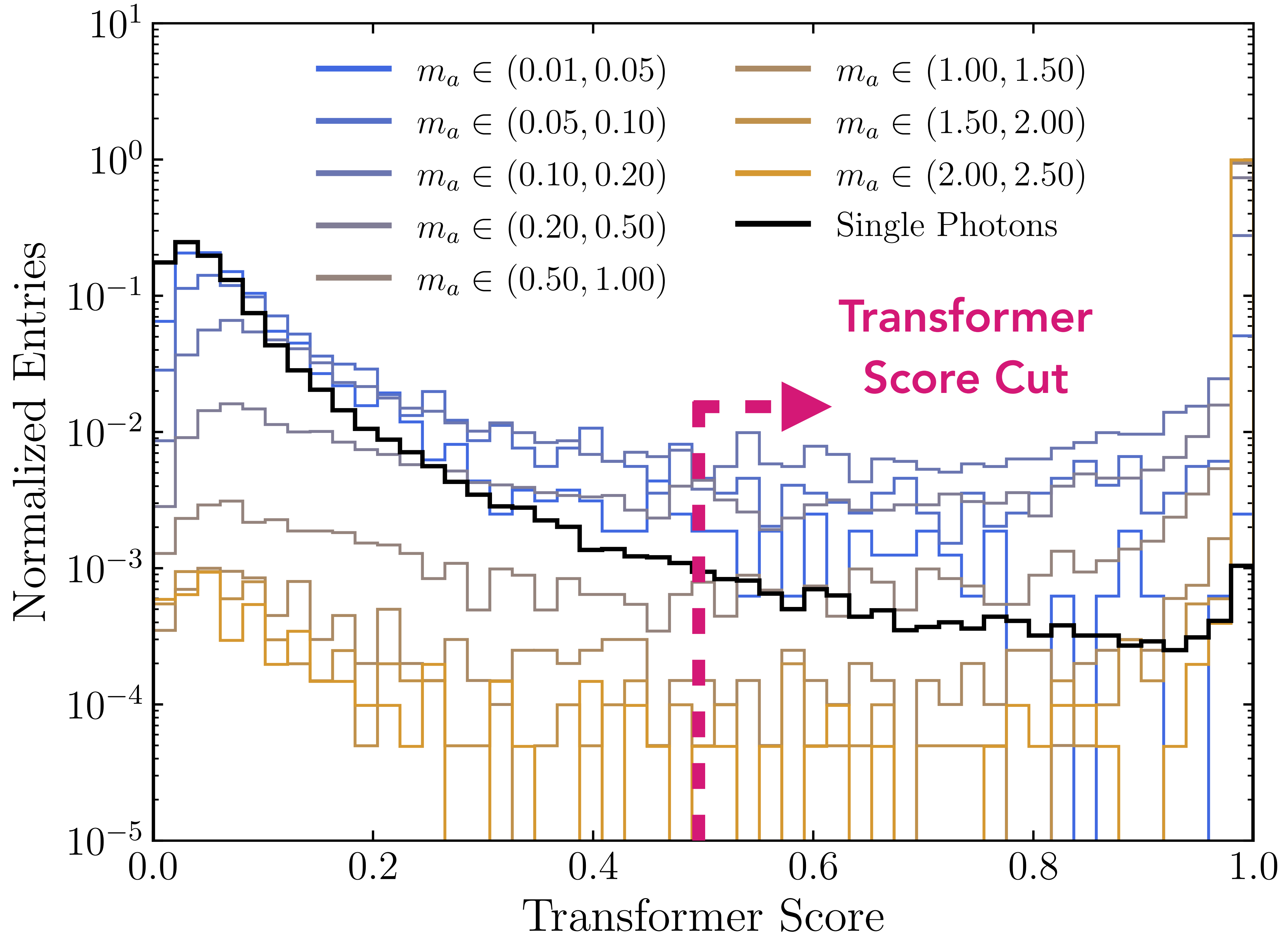
Transformer Performance: ROCs

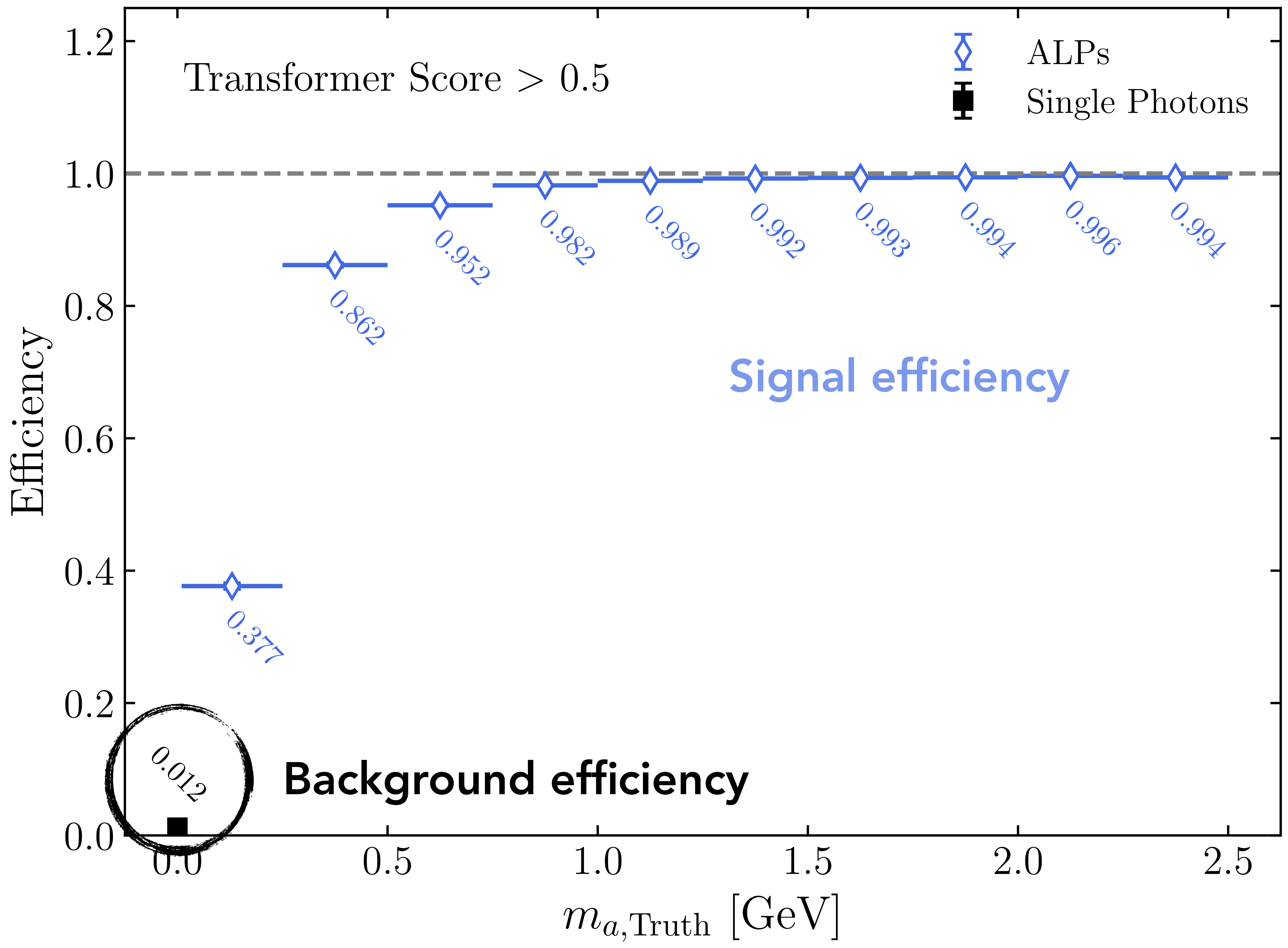
- **Excellent performance for $\Delta R_{\gamma\gamma} > 10^{-3}$ and retains discriminating power up to $\Delta R_{\gamma\gamma} \sim 3 \times 10^{-4}$**
  - The $\Delta R_{\gamma\gamma} \lesssim 10^{-3}$ region corresponds to a regime where $m_a < 100$ MeV, and the photon separation is smaller than the finest calorimeter cell in the first layer strips





Transformer Performance: Score

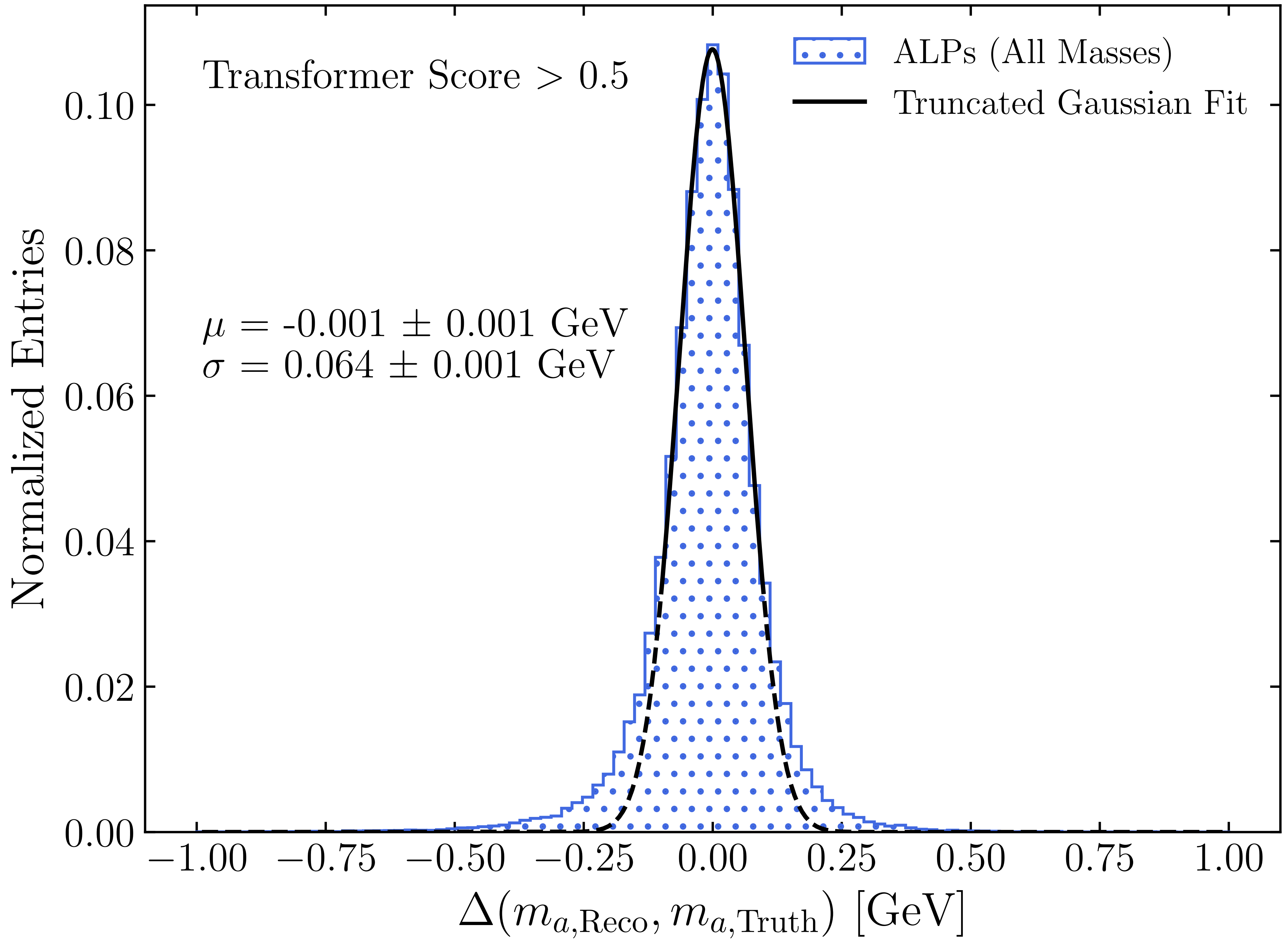


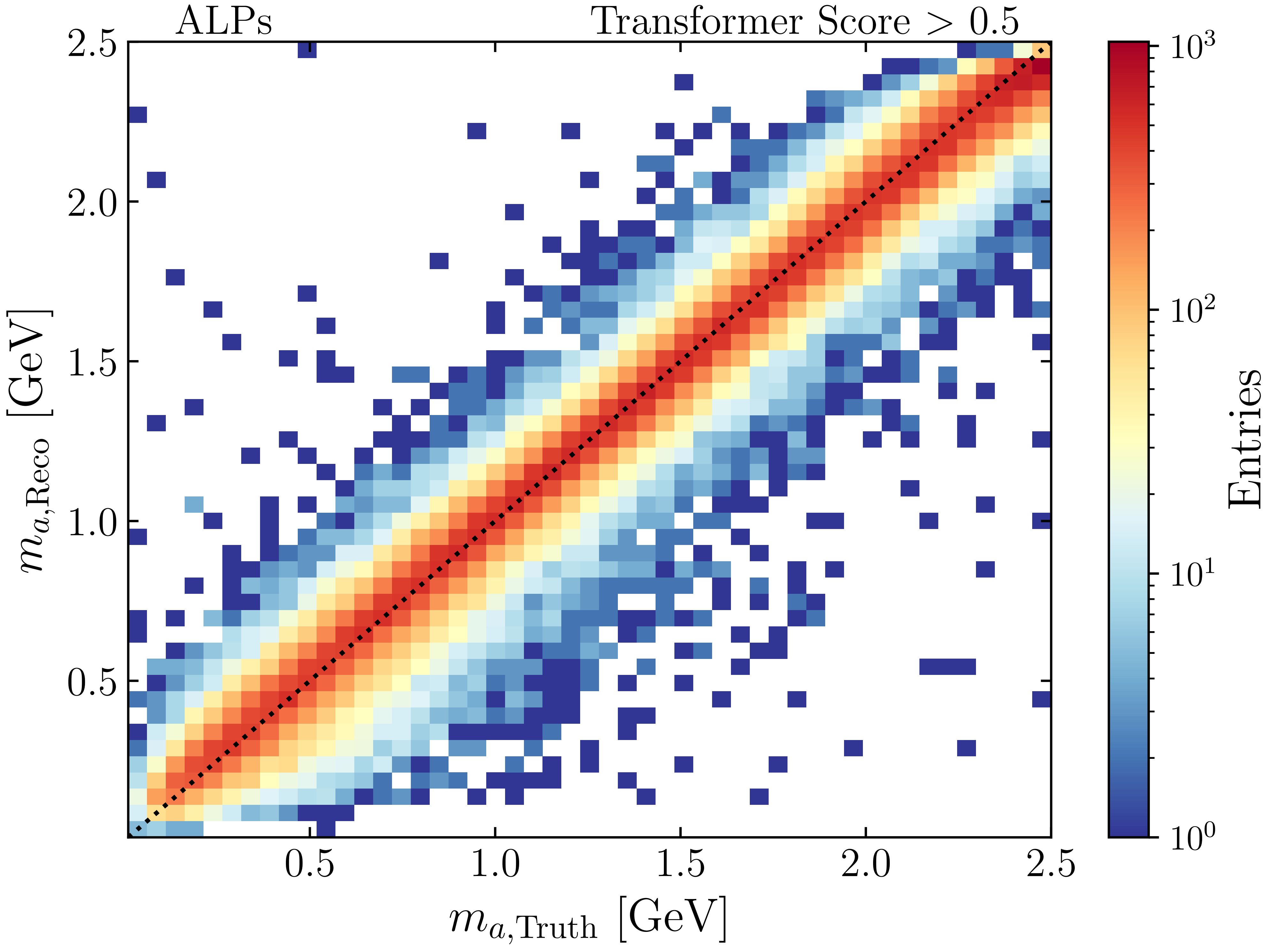


- Illustrative score selection of > 0.5 used to select good PJ candidates
  - **Signal efficiency > 95% for $m_a > 500$ MeV and background efficiency of ~1%**

Transformer Performance: Mass Regression

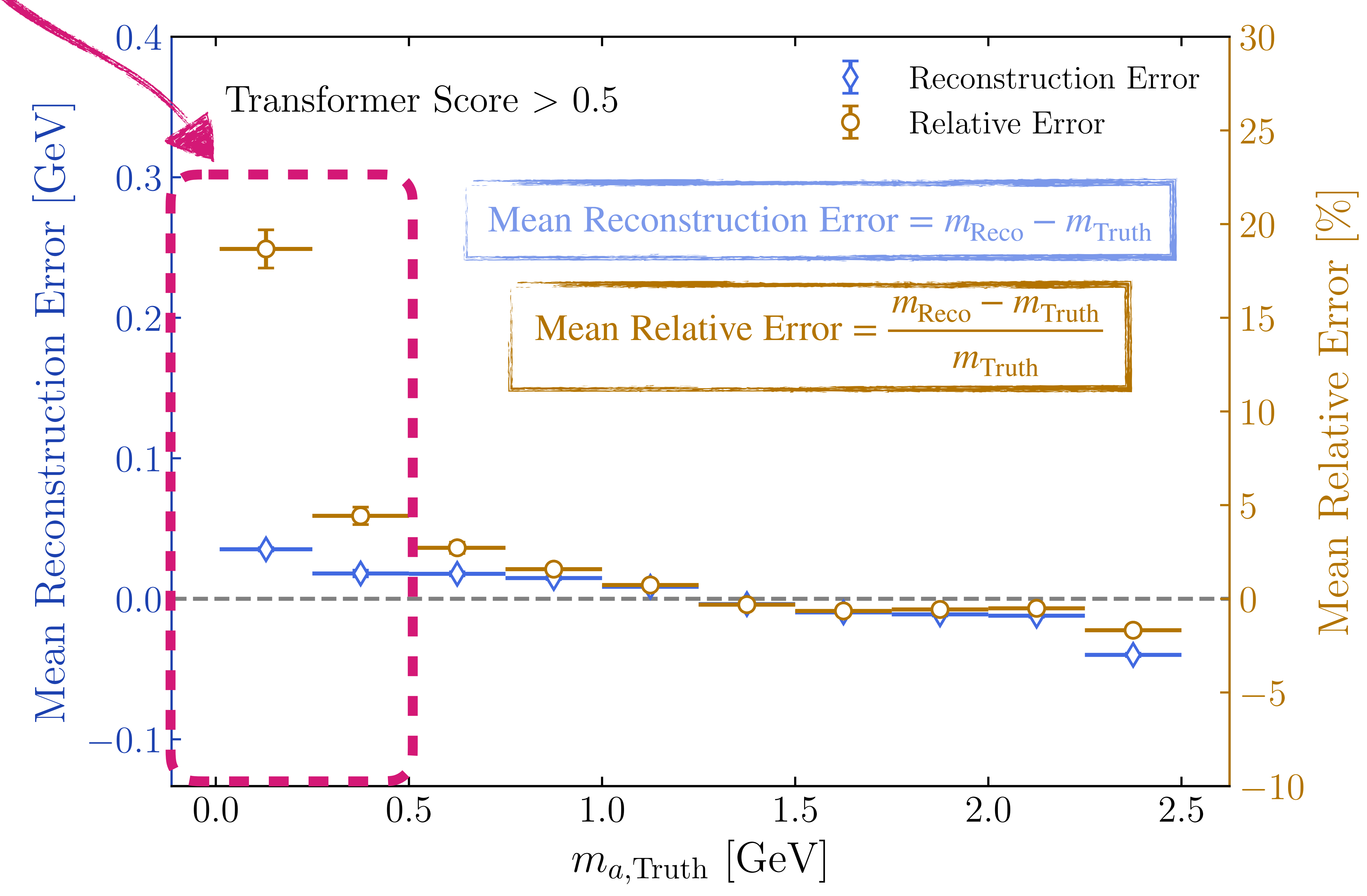
- The mass regression performance determined after Transformer score PJ quality cut
  - Accuracy (resolution) of the regression calculated from the μ (σ) of a Gaussian fit to the reco. error
  - **Overall mass resolution $\sigma_m = 64 \pm 1$ MeV with good linearity on regressed values w.r.t. $m_{a,\mathrm{Truth}}$**

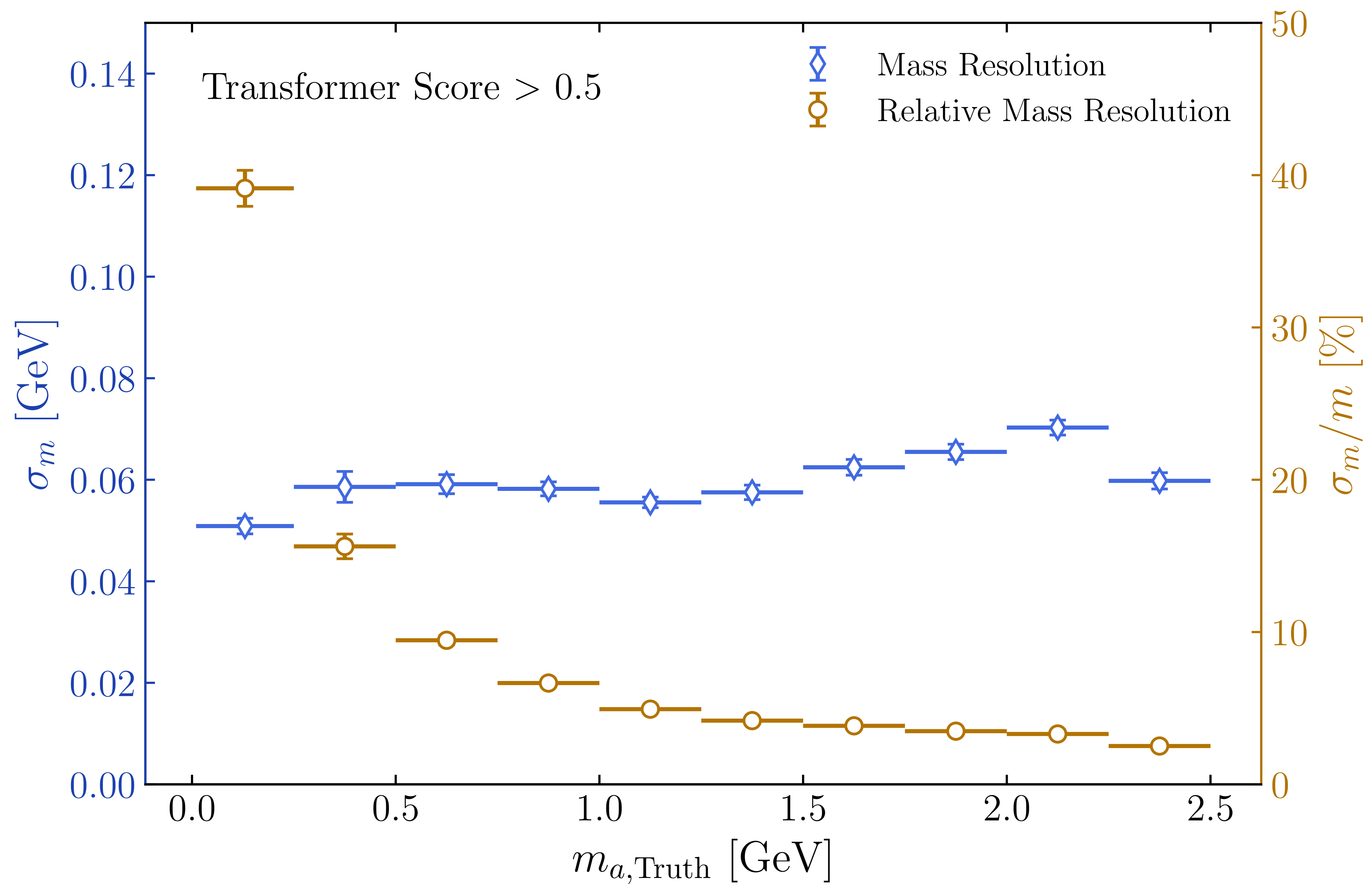




Transformer Performance: Mass Regression

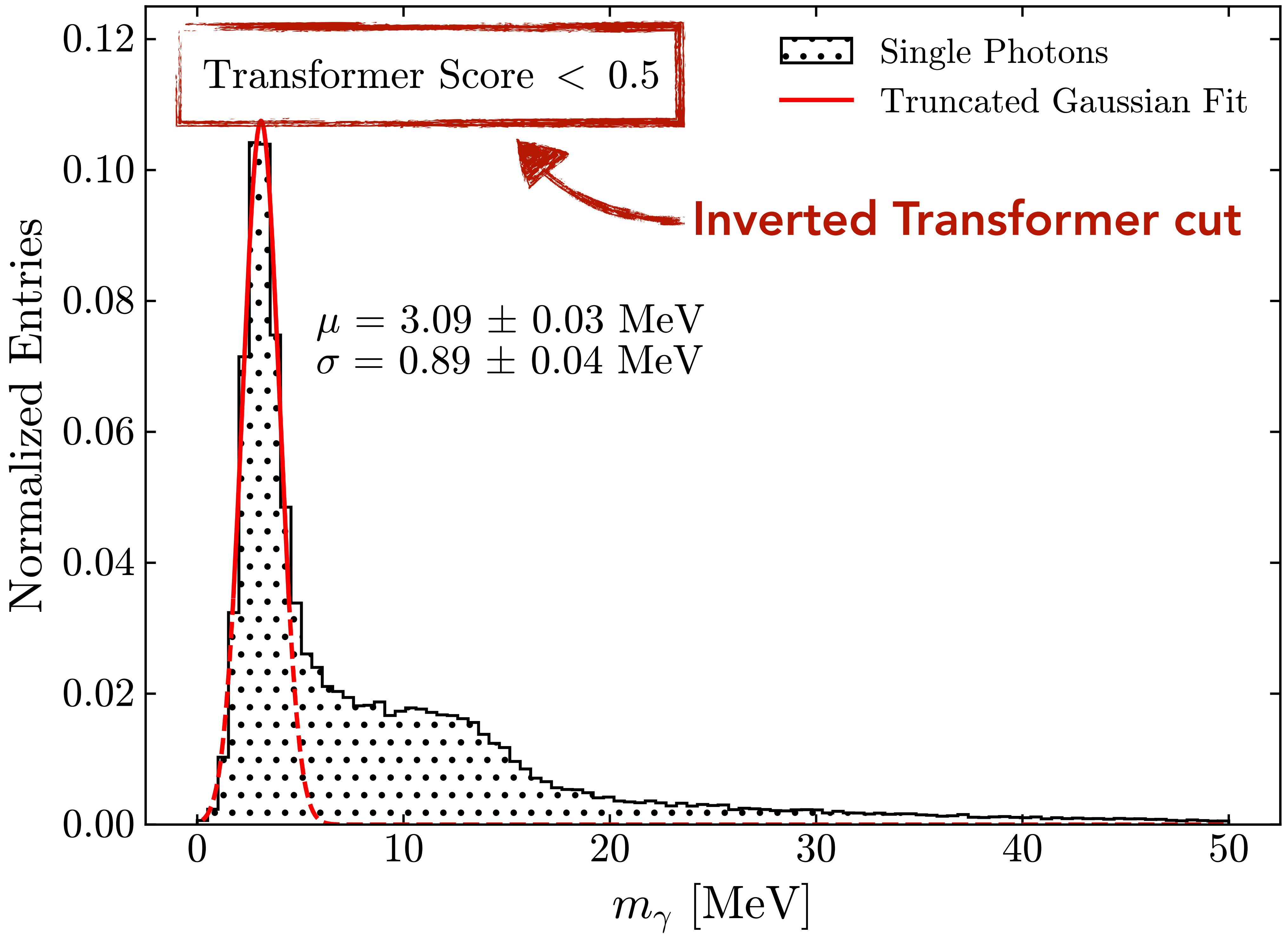
- **Both mean reconstruction error and relative mass resolution σ_m/m <5% for most m_a values**
  - Distortion at very small ALP masses due to positive-definiteness of the regression target
  - Can be addressed by artificially extending masses to negative values via **domain continuation** [link]

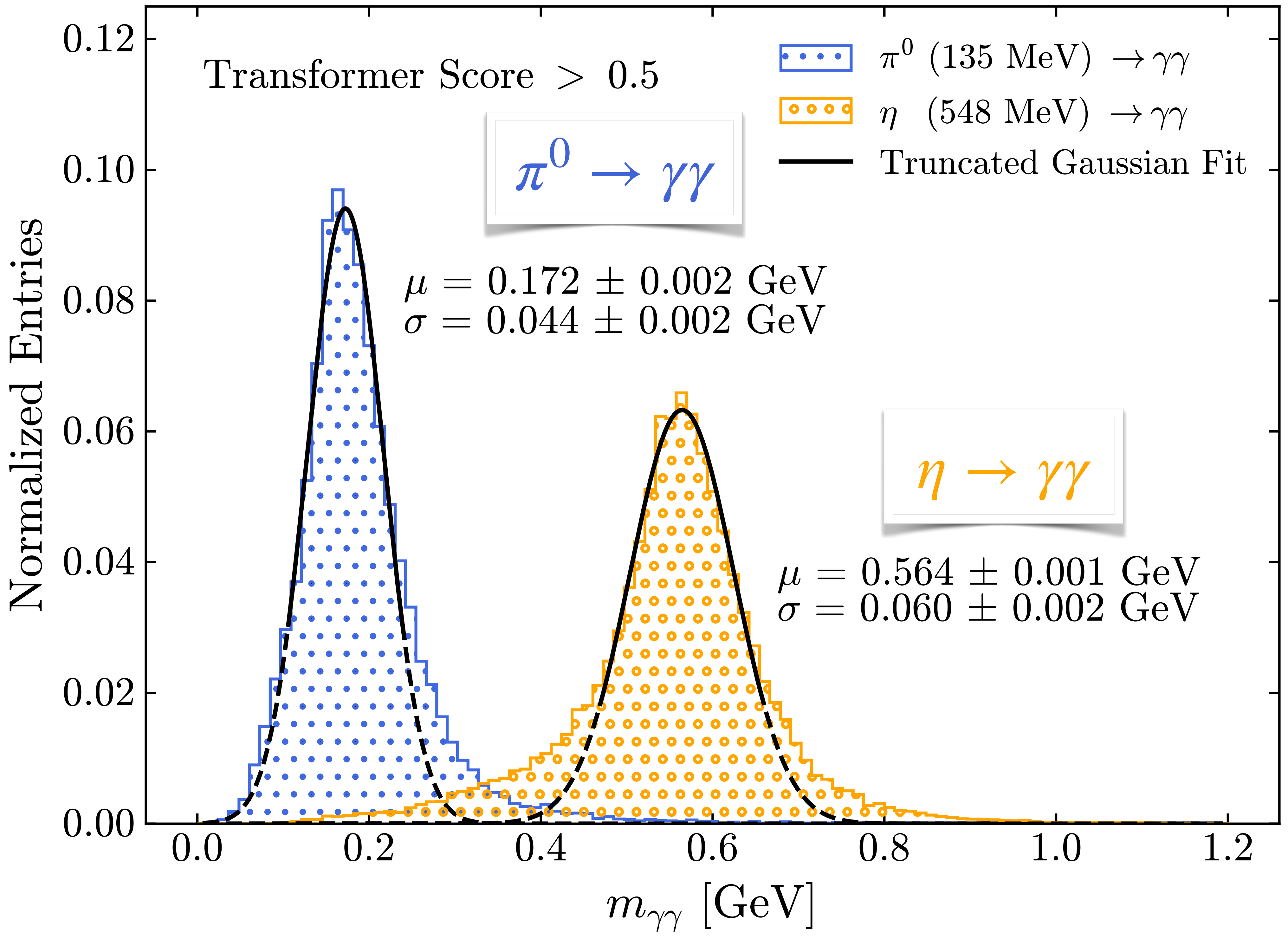




Transformer Performance: Mass Regression

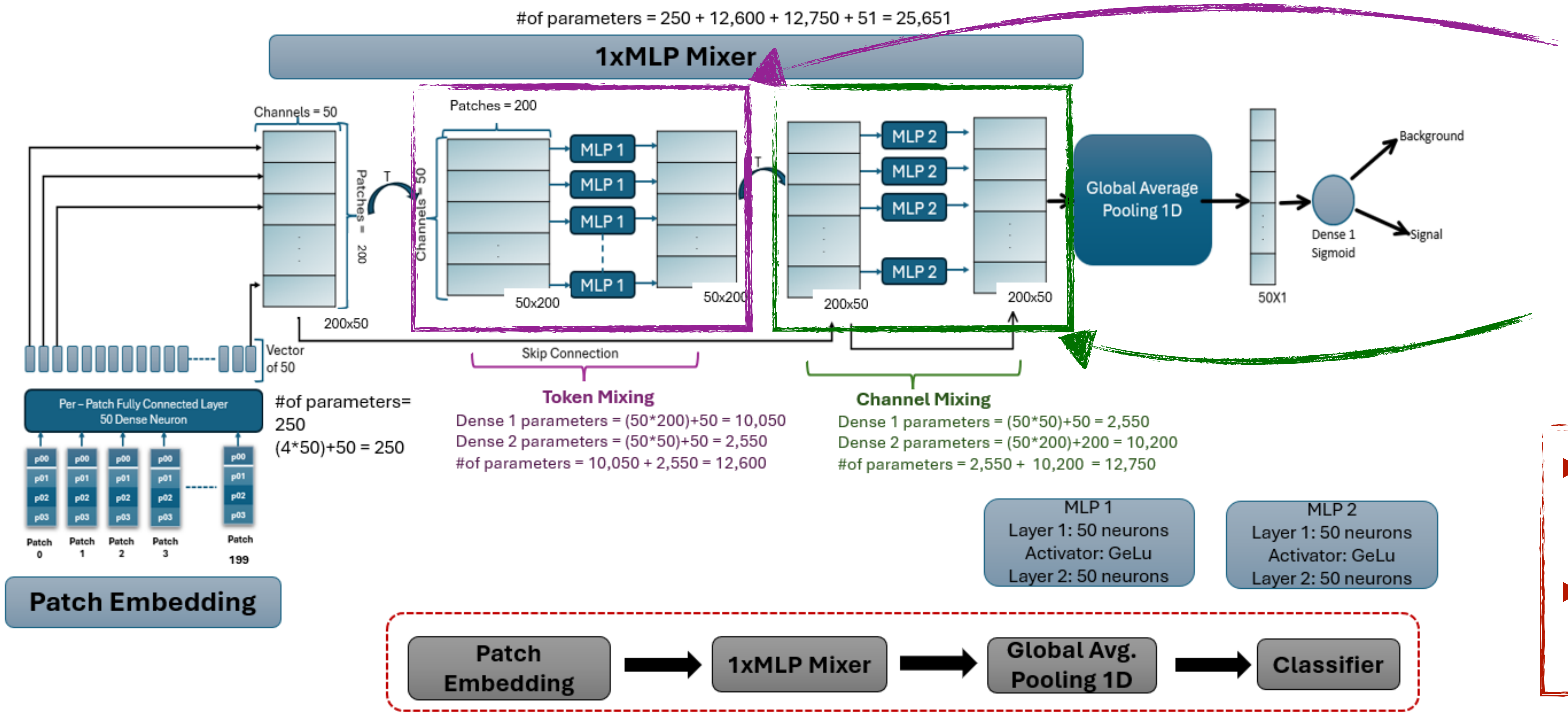
- Mass regression validation done with single photons and with an independent $\pi^0/\eta \to \gamma\gamma$ sample
  - **Transformer mass regression discriminates between π^0s and ηs, and predicts $m_\gamma \sim 3$ MeV**
  - **Provides additional handle on rejecting background and fakes**





MLP Mixer

- Sensitivity to overlapping photons in collider experiments also dependent on real-time selections
  - Transformers suffer from $\mathcal{O}(N^2)$ scaling which make them intractable for trigger deployment
  - Considered **MLP Mixer** as a lightweight Transformer alternative for potential real-time applications

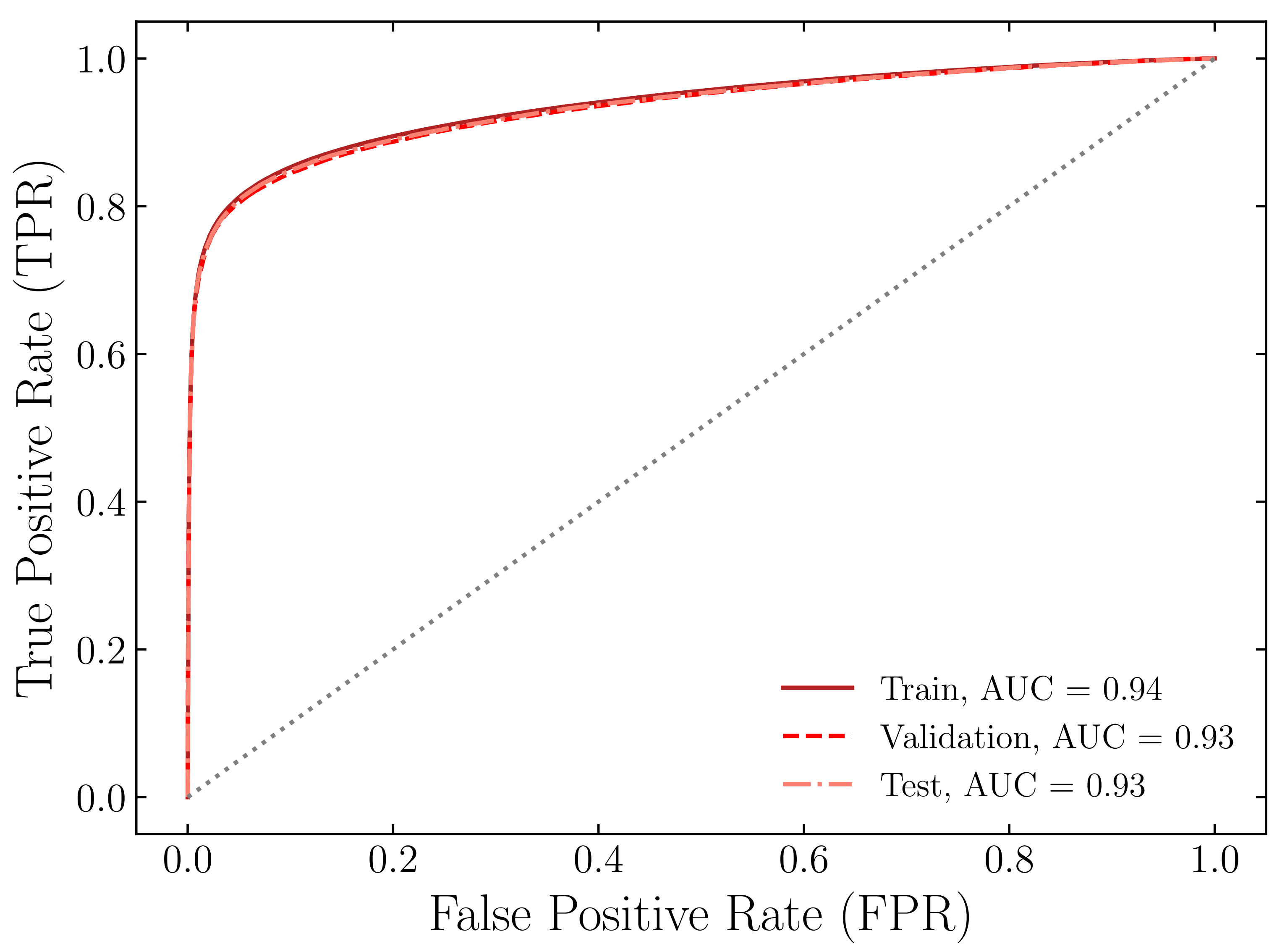


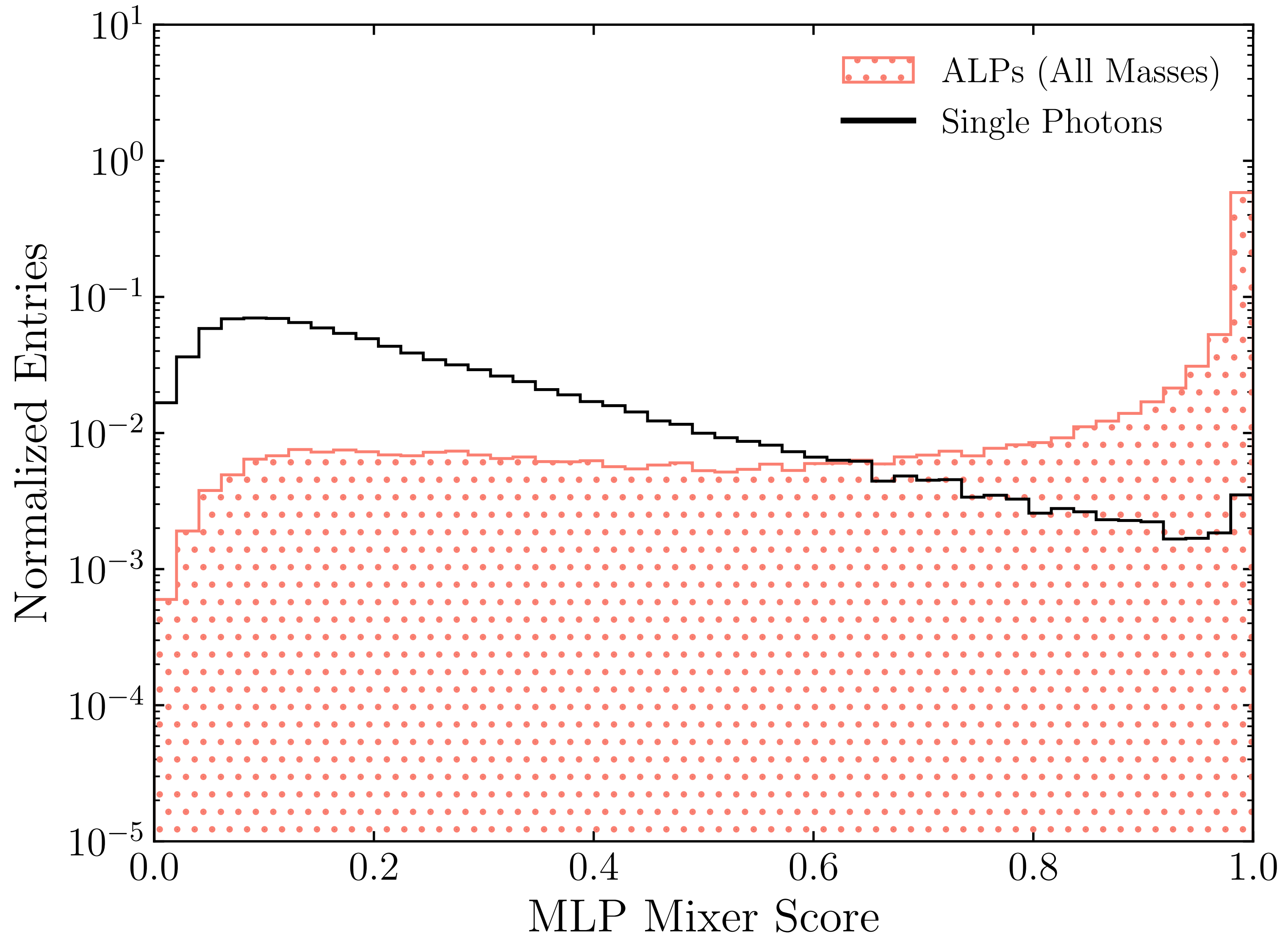
Token mixing models pairwise correlations between cells

Channel mixing models interaction between cell features

- Approximates Transformer attention mechanism with MLPs
- **Significantly more lightweight in parameter count footprint!**

MLP Mixer Performance: ROCs and Scores

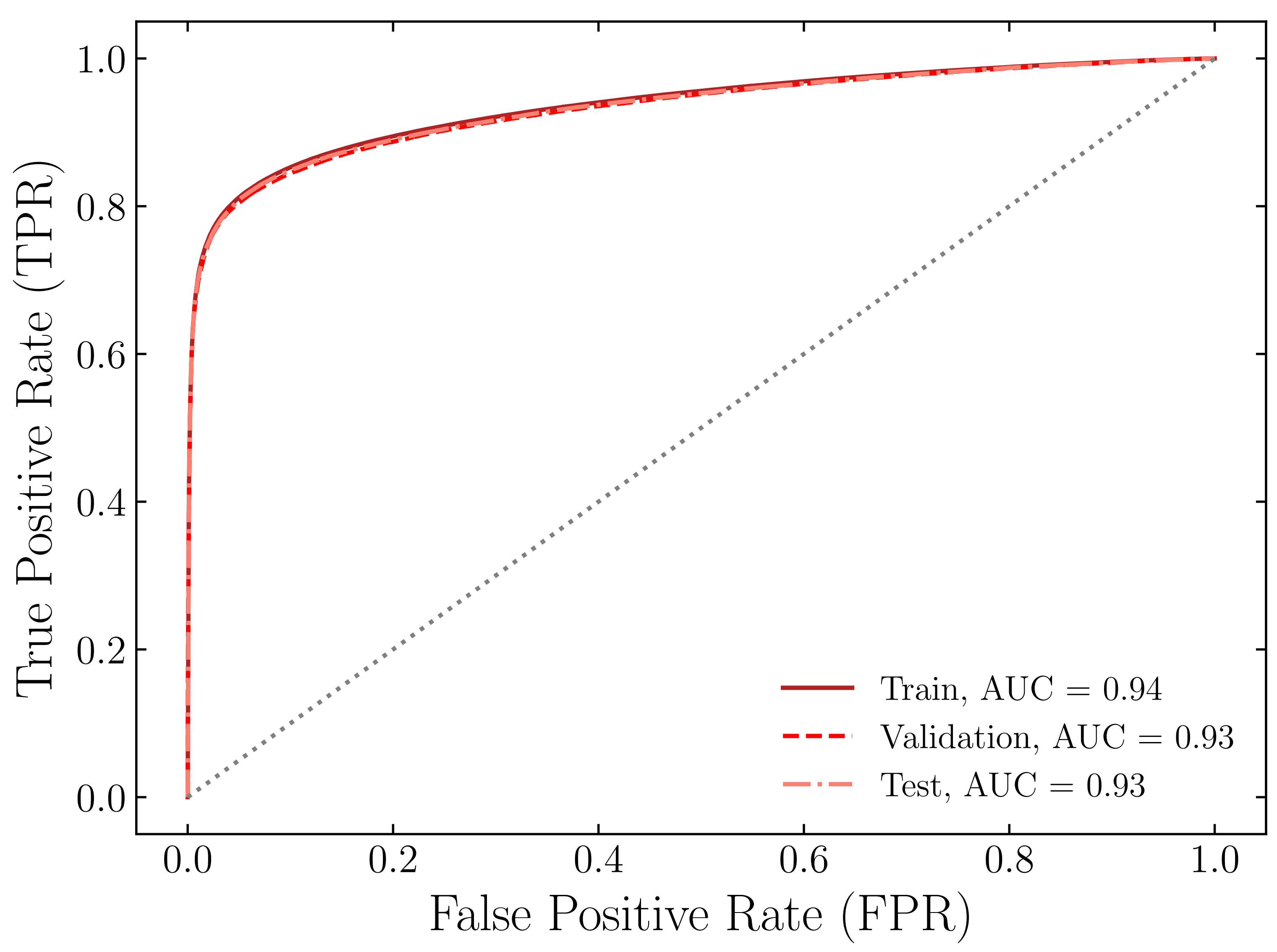


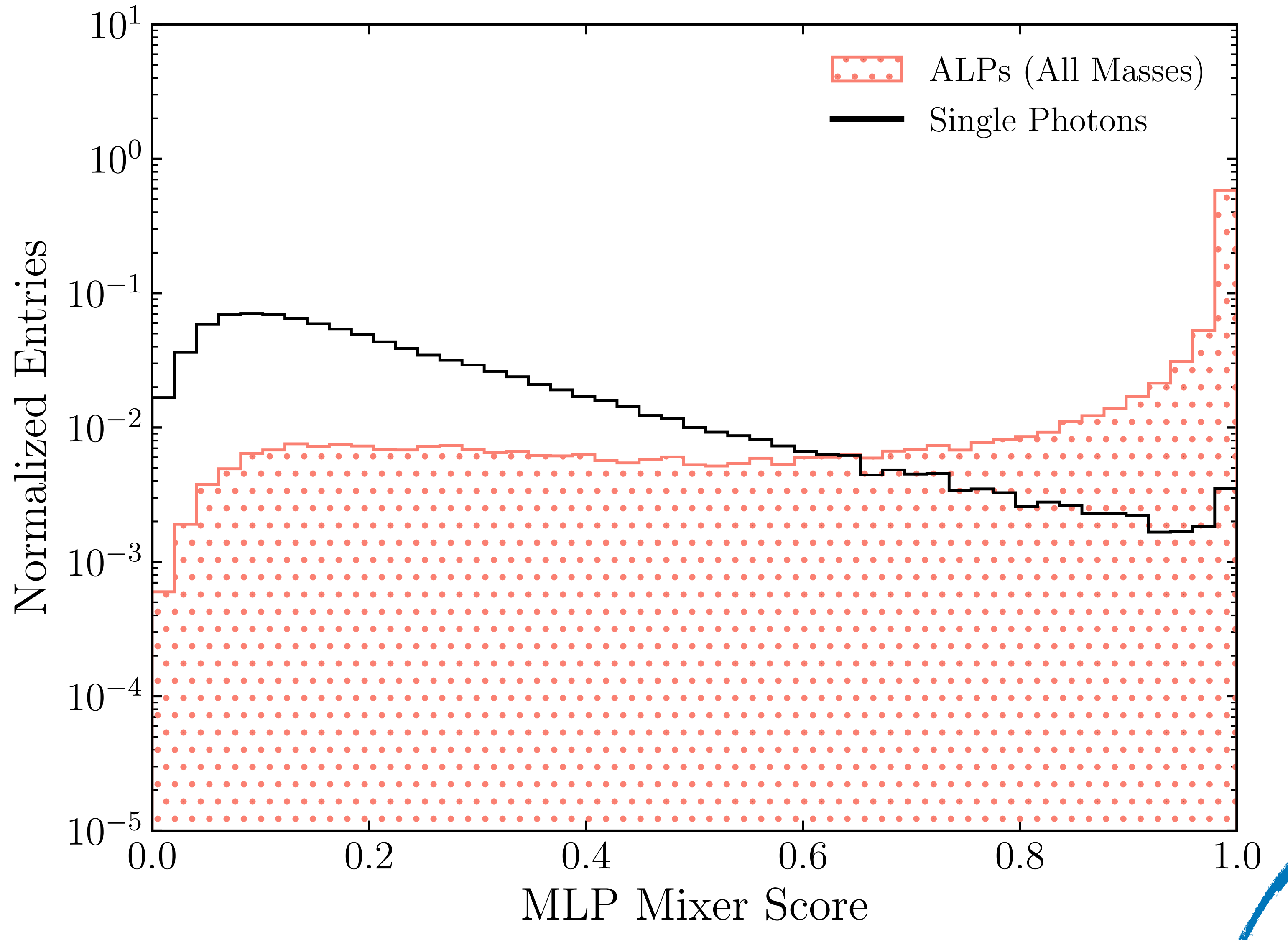


- MLP Mixer trained for 200 epochs on the benchmark dataset for PJ classification*
- Attains inclusive AUC of 0.93

*Classification only

MLP Mixer Performance: ROCs and Scores





- MLP Mixer trained for 200 epochs on the benchmark dataset for PJ classification*
- Attains inclusive AUC of 0.93

*Classification only

Floating point operations (FLOPs)

Multiply and accumulate (MAC) operations (~1/2 of FLOPs)

	Transformer	MLP Mixer
Number of Trainable Parameters	634,788	25,651
Inference FLOPs	157,083,744	4,080,000
Inference MAC Operations	78,541,872	2,040,000
Total Signal AUC	0.98	0.93

Degradation in the inclusive AUC of 5%, but using only 3% of the total Transformer FLOPs

- Potential for a trigger application to maintain low thresholds, and then apply full Transformer offline

Benchmark Models: Comparison

- Full comparison study between both SSV and cell-level approaches done with benchmark dataset

Model	Overall AUC	Background Rejection ($1/\epsilon_{\text{bkg}}$) 90% ϵ_{sig}	95% ϵ_{sig}	99% ϵ_{sig}	Parameters
SSV-Level					
BDT	0.92	3.2	1.8	1.2	$\mathcal{O}(10^3)$
DNN	0.92	3.1	1.8	1.2	
Cell-Level					
CNN	0.95	6.4	2.6	1.3	$\mathcal{O}(10^5 - 10^6)$
PFN (DeepSets)	0.97	62 ± 7	5.4 ± 0.2	1.48 ± 0.01	
Transformer	**0.98**	$\mathbf{472 \pm 60}$	$\mathbf{9.0 \pm 0.4}$	$\mathbf{1.64 \pm 0.01}$	
MLP Mixer	**0.93**	**4.3**	**2.1**	**1.2**	$\mathcal{O}(10^4)$

All cell-level approaches **outperform the SSV based methods**, with the **Transformer providing x100 better background rejection*** and the **MLP Mixer a lightweight cell-level alternative**

*at 90% ϵ_{sig}

Concluding Remarks

- ML with **calorimeter cells** on a simulation of an ATLAS-like calorimeter for the identification of $a \rightarrow \gamma\gamma$ photon jets from single photons outperforms previous methods implemented using SSVs
  - Study on benchmark dataset shows that **Transformers are the most performant architecture**, with an **MLP Mixer providing a smaller footprint alternative** for a potential trigger application
- Results submitted to JHEP, with ongoing efforts to apply this methodology directly to ATLAS

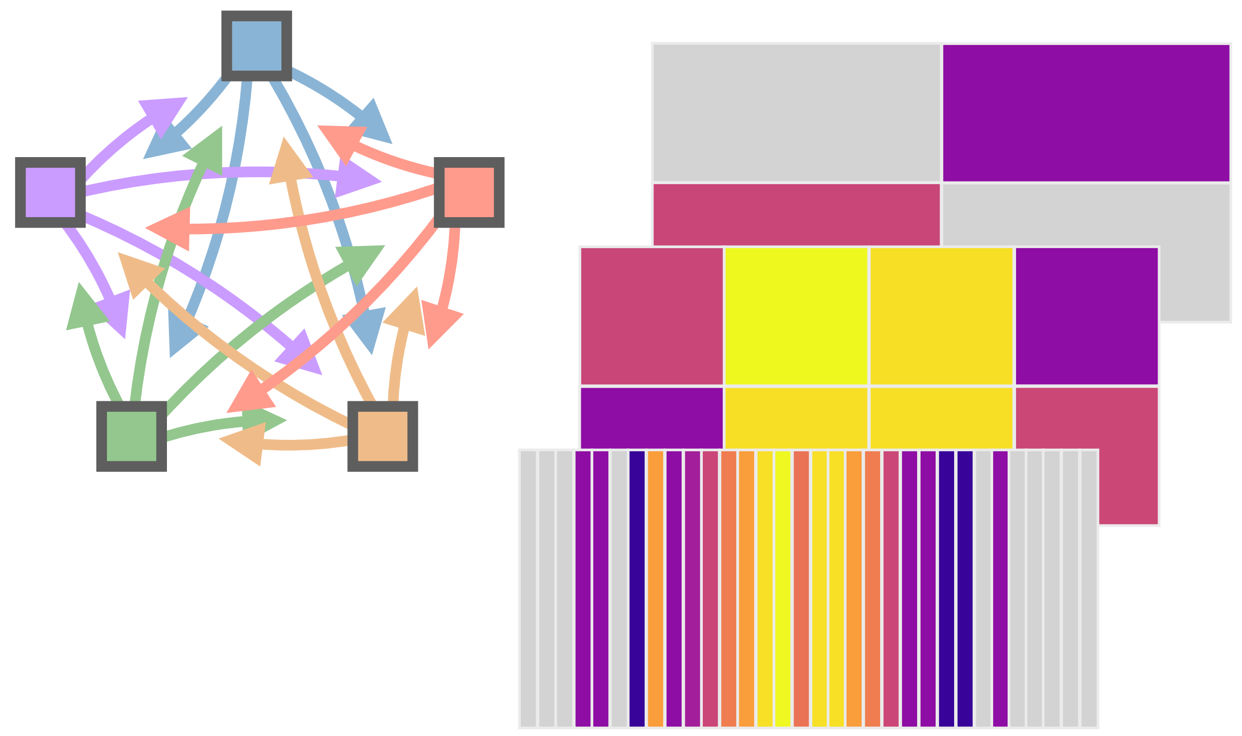
[Submitted on 9 Jul 2026]

Transformer-based machine learning using low-level calorimeter signals for collimated photon identification at collider experiments

Gabriel Matos, Lauren Larson, Abhilasha Dave, Maria Bressan, Azal Amer, Cindy Liu, Nikiforos Nikiforou, Jonathan Long, Timothy Andeen, John Parsons, Julia Gonski

Electromagnetic calorimeters provide essential information for reconstructing and selecting both Standard Model (SM) and potential beyond the SM physics events at high-energy particle colliders. The fine-grained segmentation of modern calorimeters captures rich information about the internal structure of particle showers, much of which is discarded by conventional high-level reconstruction methods. In this work, we leverage calorimeter cell-level information to classify highly collimated diphoton signatures, arising from the decay of light axion-like particles, from isolated single-photon showers. We systematically compare a range of machine learning architectures, spanning high-level, shower shape variable-based approaches and direct cell-level methods. Cell-level machine learning shows significantly superior classification ability, with a Transformer in particular representing the best performance among six different architectures studied, and an MLP Mixer representing a resource-constrained alternative for potential real-time, trigger-level applications. Beyond classification, the Transformer model developed enables direct invariant mass regression from calorimeter cells, improving the characterization of light resonances and providing an additional handle in reducing the π^0 and η fake photon backgrounds. These results demonstrate that cell-level machine learning methods can extend calorimeter-based particle identification and performance well beyond the capabilities of current conventional techniques.

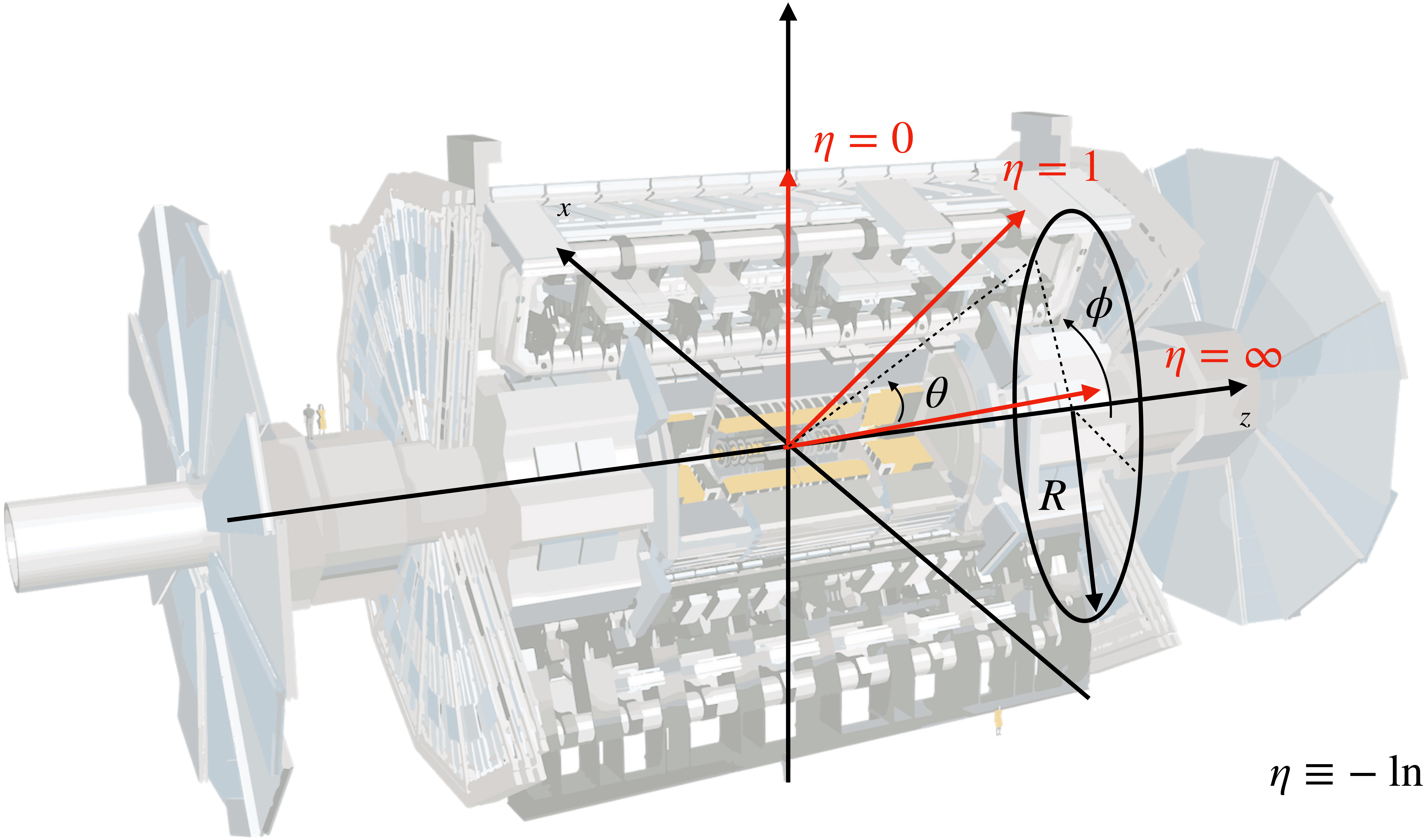
[arXiv:2607.08175]



Thanks!

Backup

ATLAS Detector Coordinates



$$\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]$$

ATLAS Topological Clustering

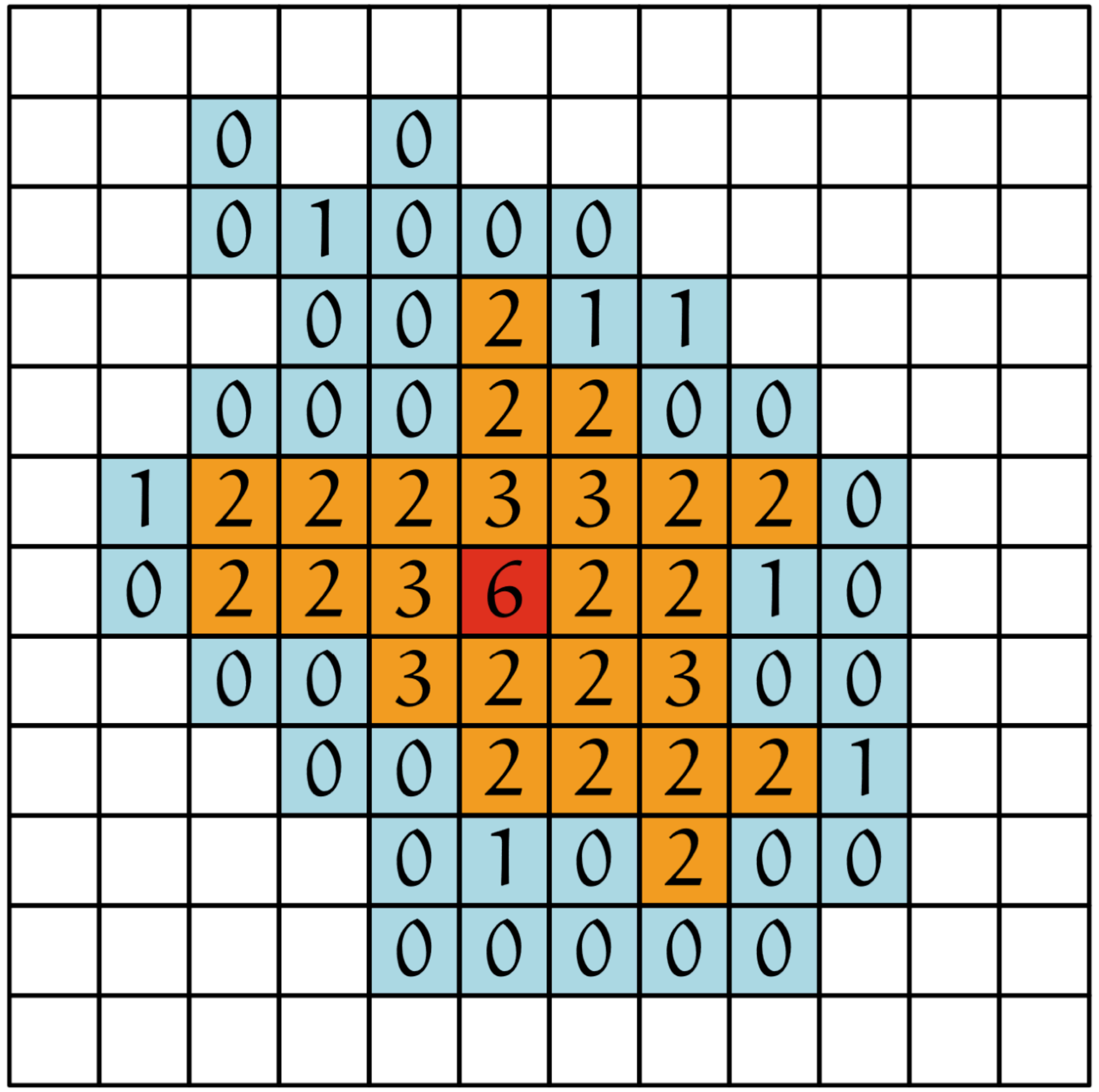


Image source [link]

- Topological clusters formed dynamically in η, ϕ via 4-2-0 cell significance prescription

$$\boxed{\zeta^{\mathrm{EM}}_{\mathrm{cell}} = \frac{E^{\mathrm{EM}}_{\mathrm{cell}}}{\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}}}$$

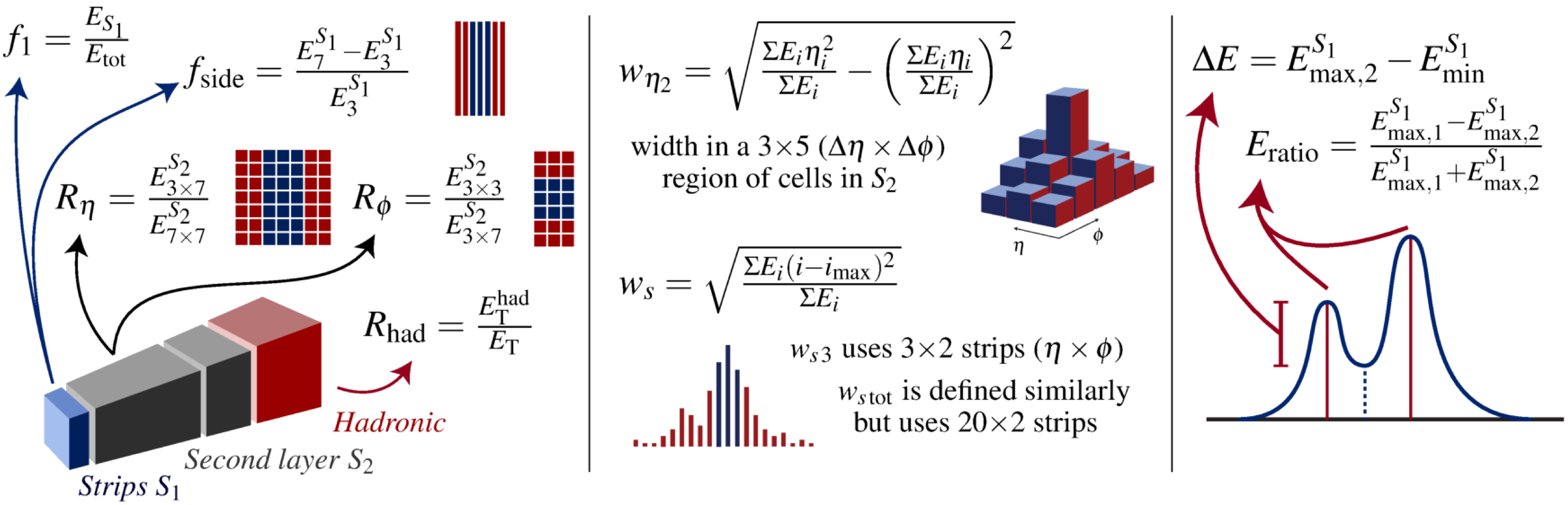
$\zeta^{\mathrm{EM}}_{\mathrm{cell}} > 4$: seed cells

$\zeta^{\mathrm{EM}}_{\mathrm{cell}} > 2$: growth cells

$\zeta^{\mathrm{EM}}_{\mathrm{cell}} > 0$: boundary cells

- As no topoclustering is applied to the ATLAS-like benchmark data, $n_{\mathrm{cells}} = 200$ is chosen to roughly match the cell multiplicity with some safety margin

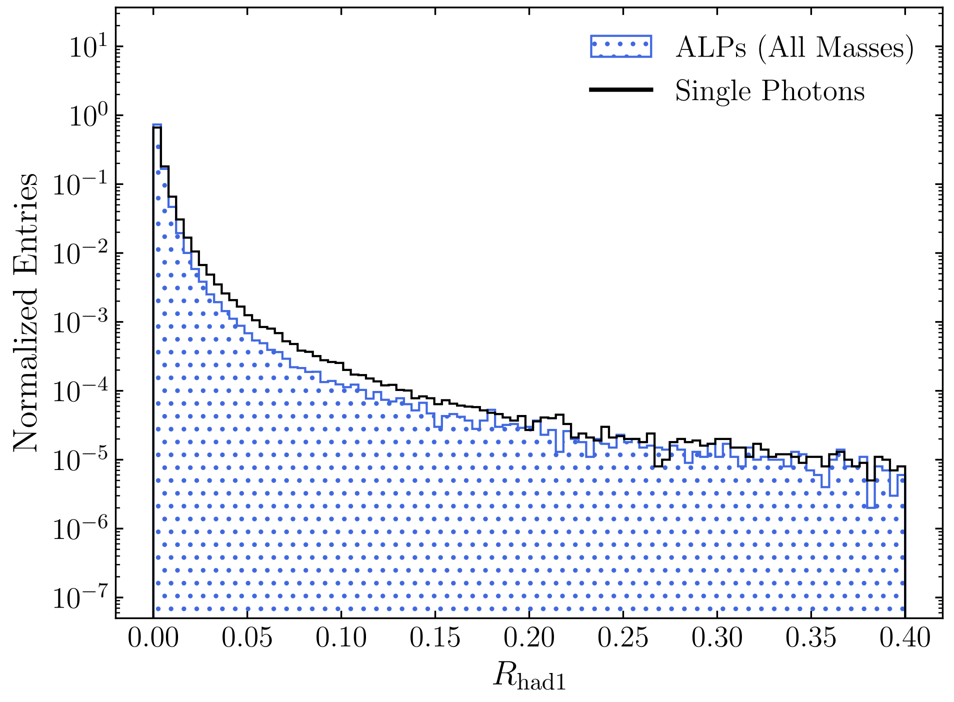
Shower Shape Variables (1)

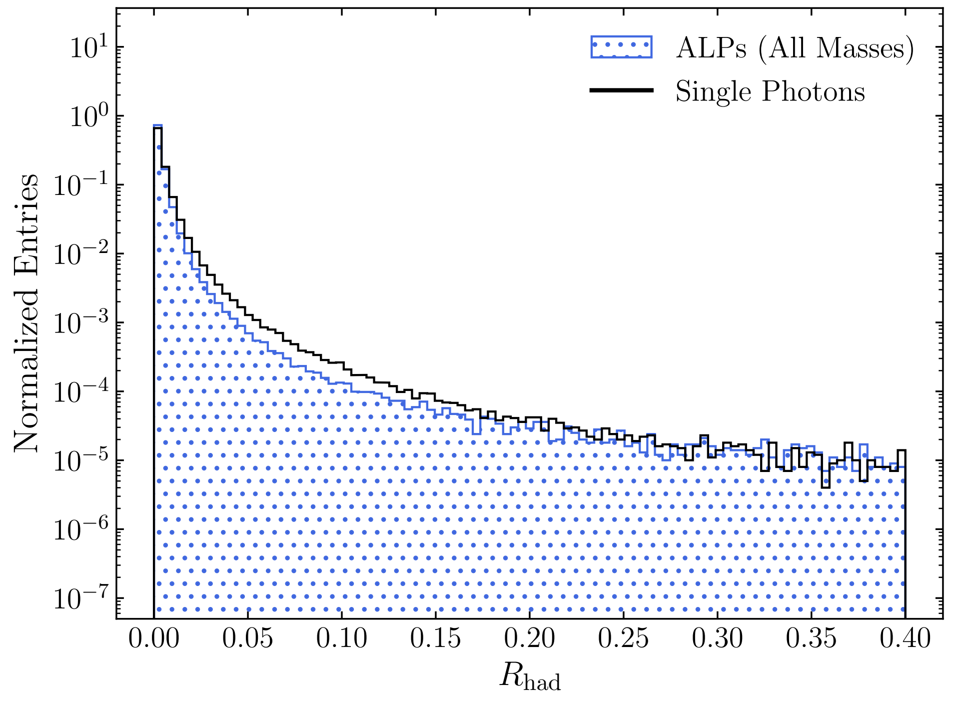


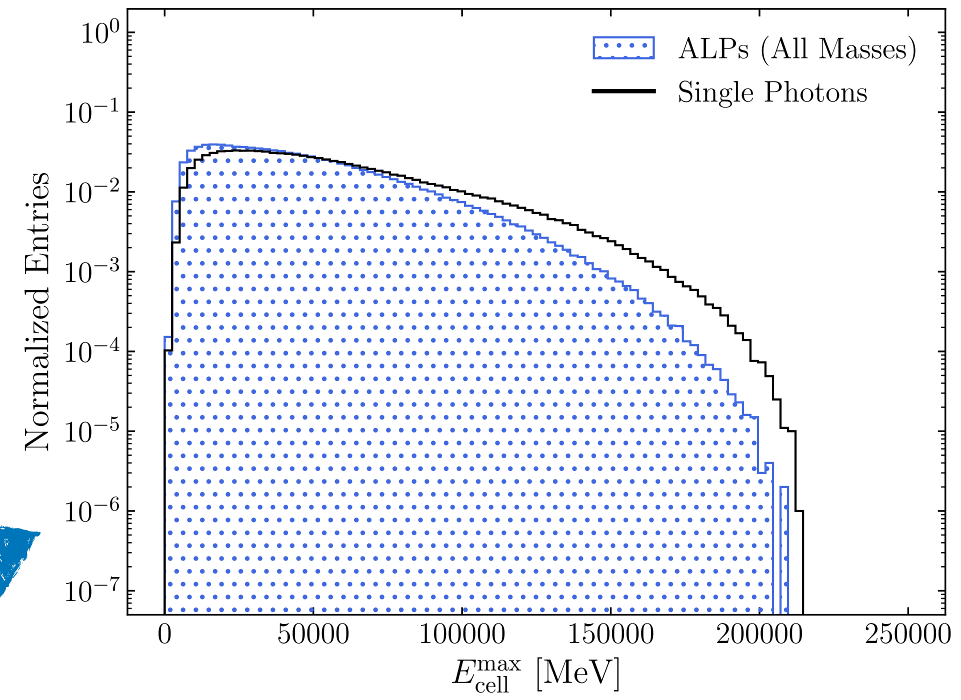
Shower Shape Variables (2)

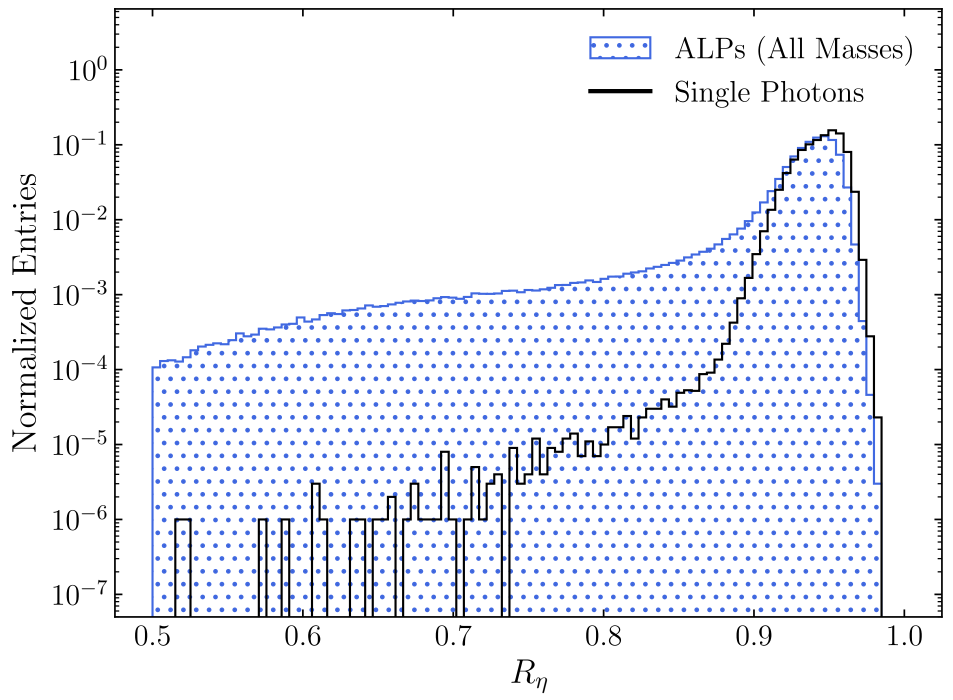
Hadronic leakage	Ratio of E_T in the first sampling layer of the hadronic calorimeter to E_T of the EM cluster (used over the range $\|\eta\| < 0.8$ or $\|\eta\| > 1.52$)	R_{had_1}
	Ratio of E_T in the hadronic calorimeter to E_T of the EM cluster (used over the range $0.8 < \|\eta\| < 1.37$)	R_{had}

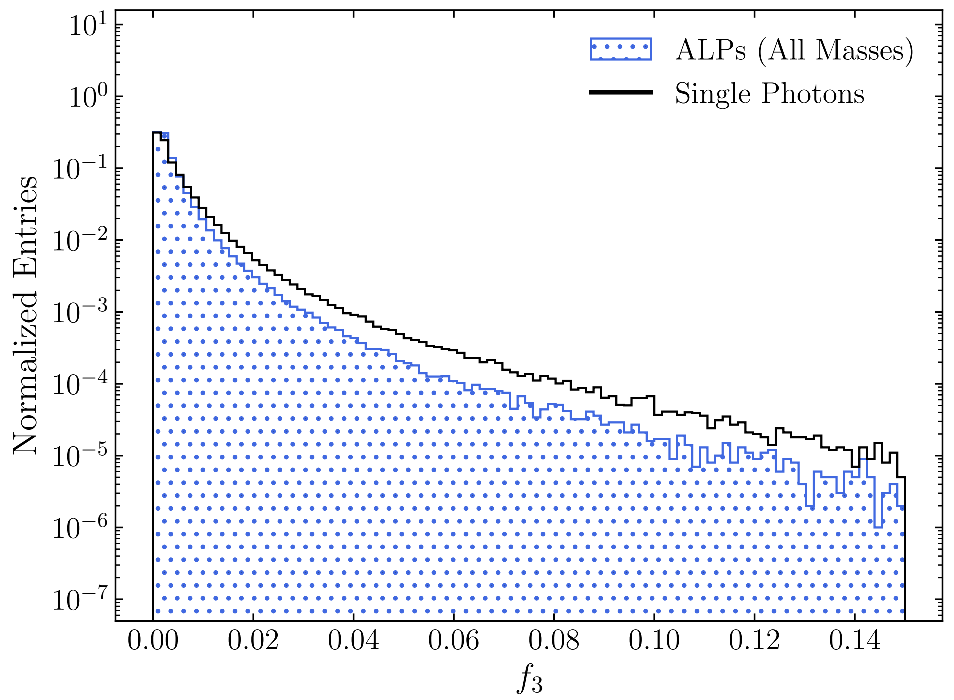
EM middle layer	Ratio of the energy in 3×7 $\eta \times \phi$ cells over the energy in 7×7 cells centered around the photon cluster position	R_η
	Lateral shower width, $\sqrt{(\Sigma E_i \eta_i^2)/(\Sigma E_i) - ((\Sigma E_i \eta_i)/(\Sigma E_i))^2}$, where E_i is the energy and η_i is the pseudorapidity of cell i and the sum is calculated within a window of 3×5 cells	w_{η_2}
	Ratio of the energy in 3×3 $\eta \times \phi$ cells over the energy of 3×7 cells centered around the photon cluster position	R_ϕ





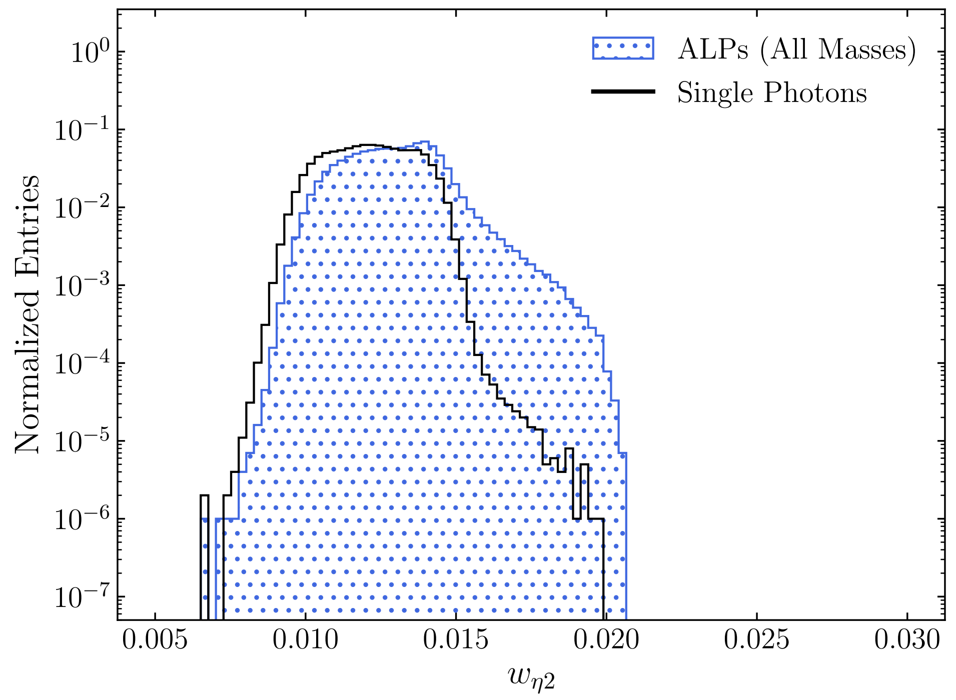


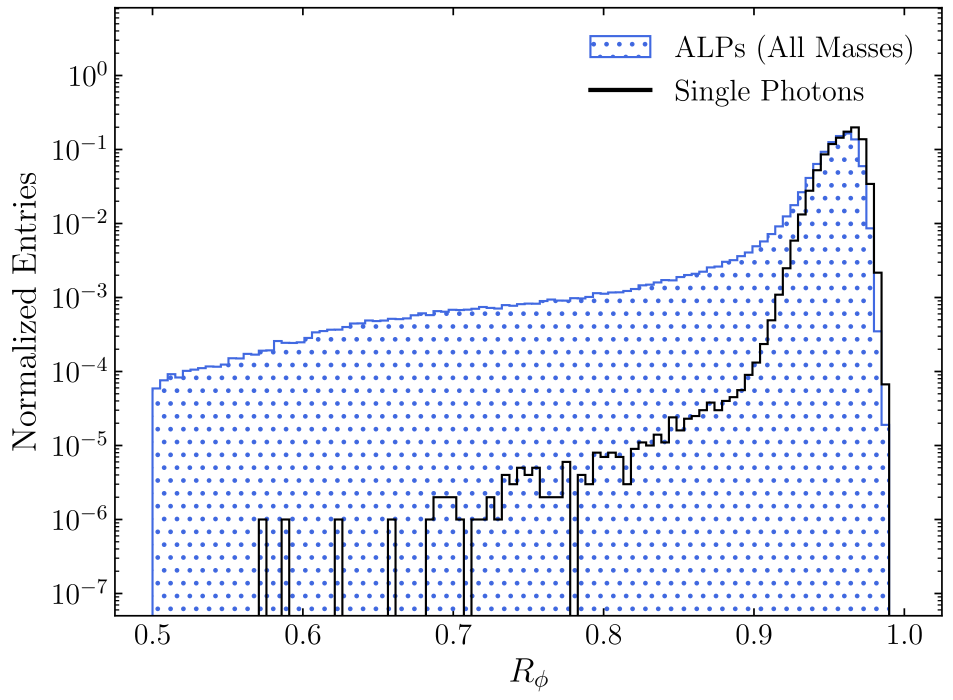




Maximum energy cell in middle layer, E_{cell}^{max}

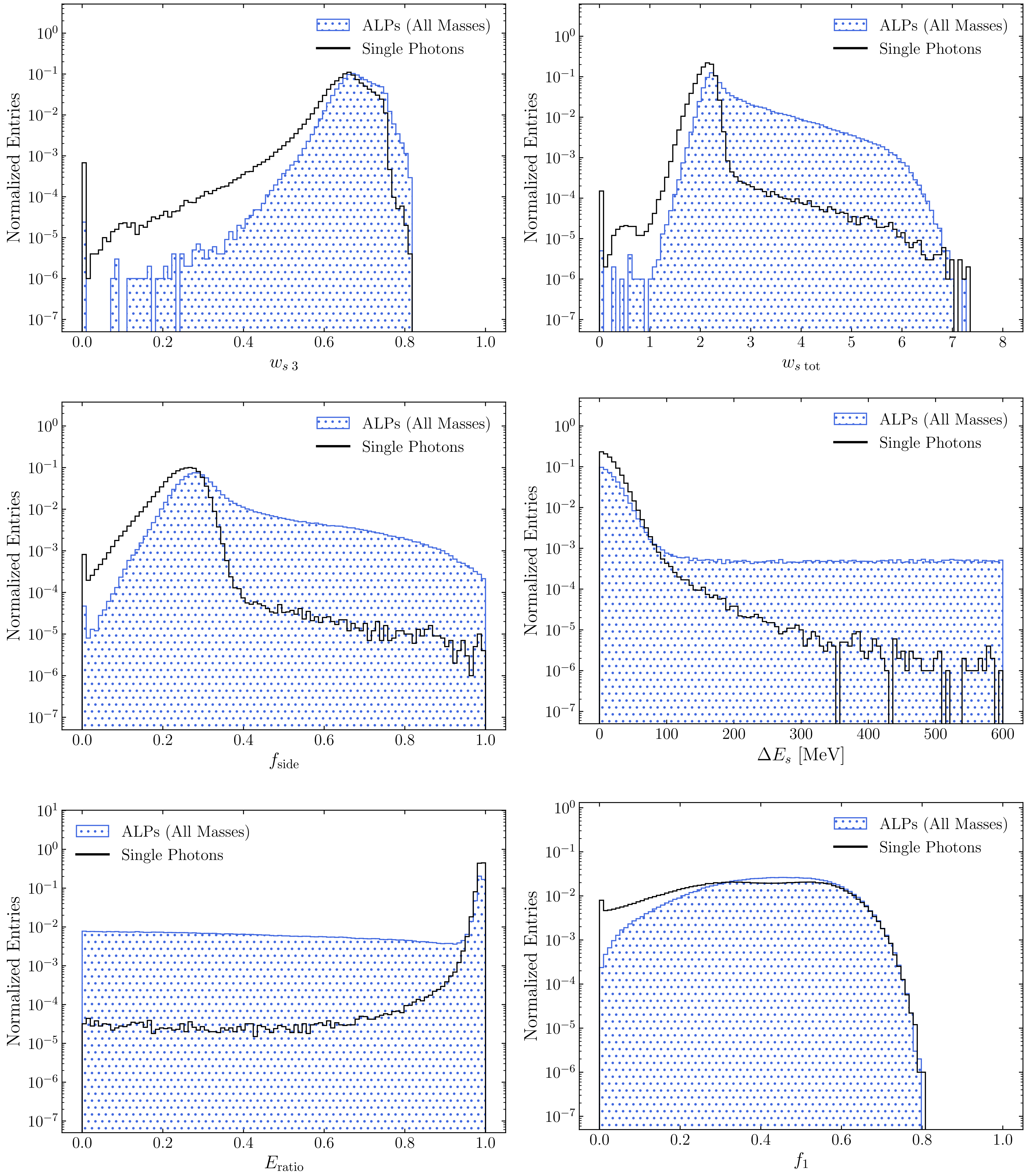
Ratio $f_3 = E_{S3}/E_{tot}$





Shower Shape Variables (3)

EM strip layer	Lateral shower width, $\sqrt{(\Sigma E_i(i-i_{\max})^2)/(\Sigma E_i)}$, where i runs over all strips in a window of 3×2 $\eta \times \phi$ strips, and $i_{\max}$ is the index of the highest-energy strip calculated from three strips around the strip with maximum energy deposit	$w_{s\,3}$
	Total lateral shower width $\sqrt{(\Sigma E_i(i-i_{\max})^2)/(\Sigma E_i)}$, where i runs over all strips in a window of 20×2 $\eta \times \phi$ strips, and $i_{\max}$ is the index of the highest-energy strip measured in the strip layer	$w_{s\,\mathrm{tot}}$
	Energy outside the core of the three central strips but within seven strips divided by energy within the three central strips	f_{side}
	Difference between the energy associated with the second maximum in the strip layer and the energy reconstructed in the strip with the minimum value found between the first and second maxima	ΔE_s
	Ratio of the energy difference between the maximum energy deposit and the energy deposit in the secondary maximum in the cluster to the sum of these energies	E_{ratio}
	Ratio of the energy in the first layer to the to the total energy of the EM cluster	f_1



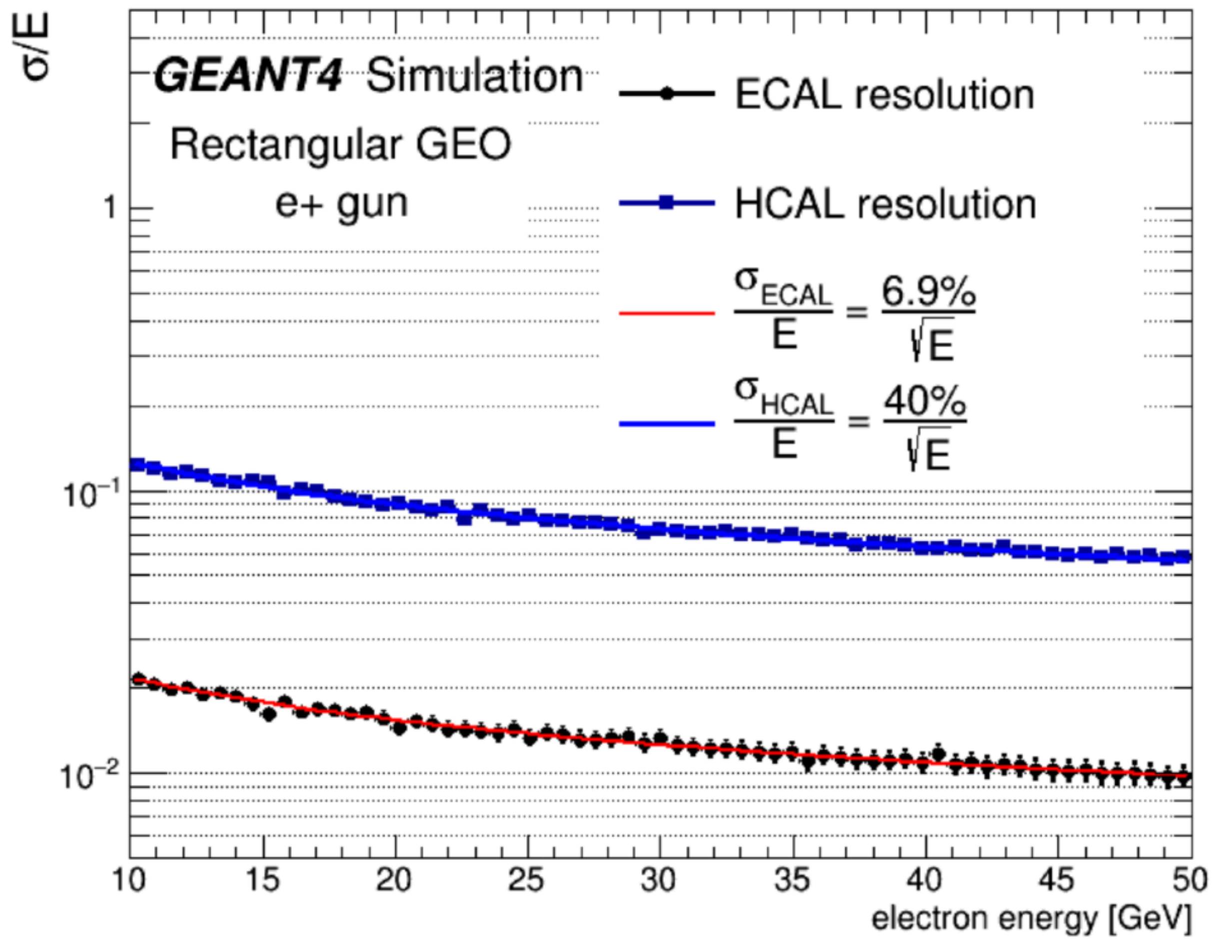
ATLAS-like Calorimeter: Details

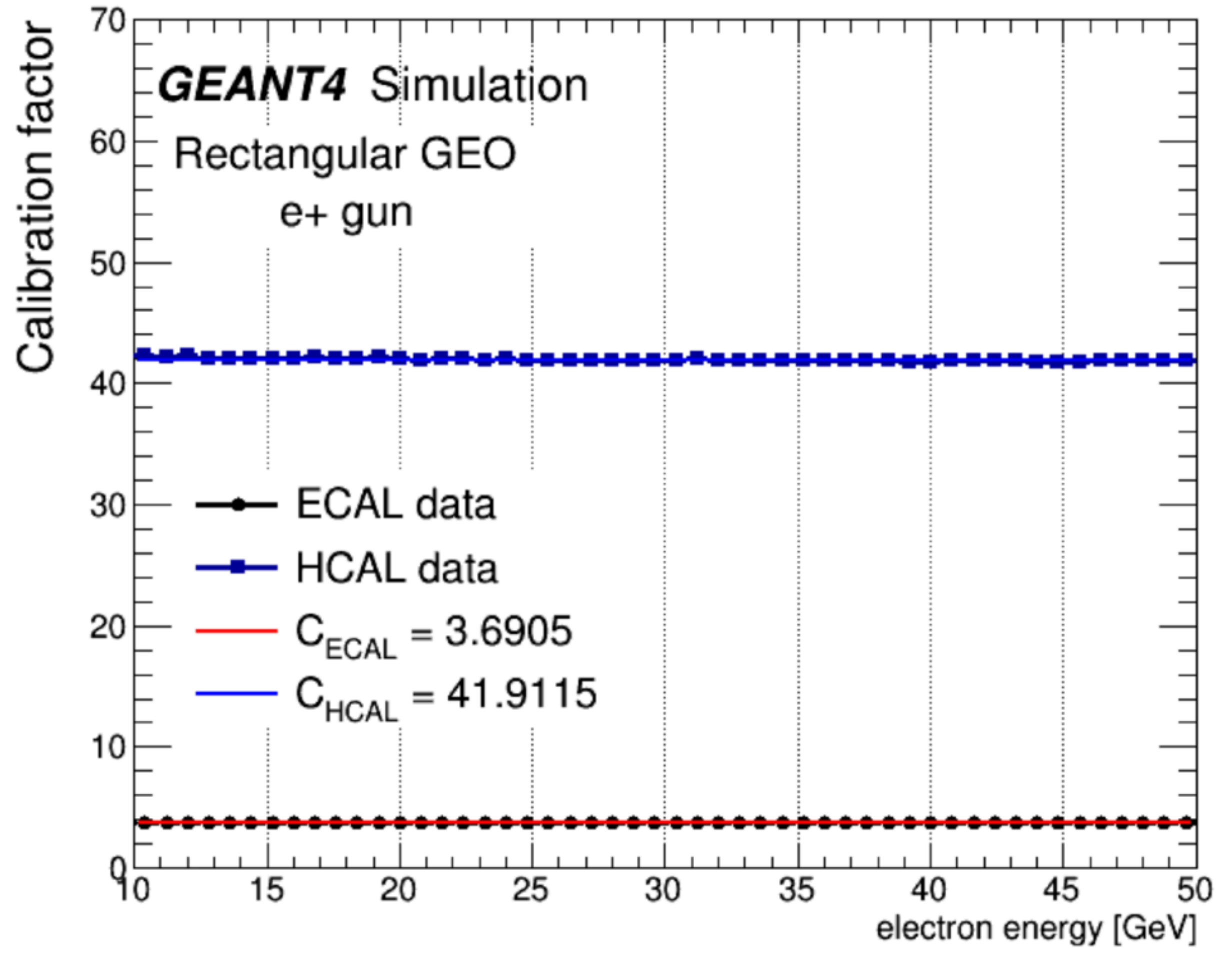
*GitHub: g4-atlas-calorimeter

- "Sandwich" calorimeter with flat faces, particle gun incidence perpendicular to calorimeter*
- Granularity of model designed to match the ATLAS calorimeter at $\eta = 0$, with 64 x 64 ECAL2 cells
- Idealized, no noise simulation, no topoclustering

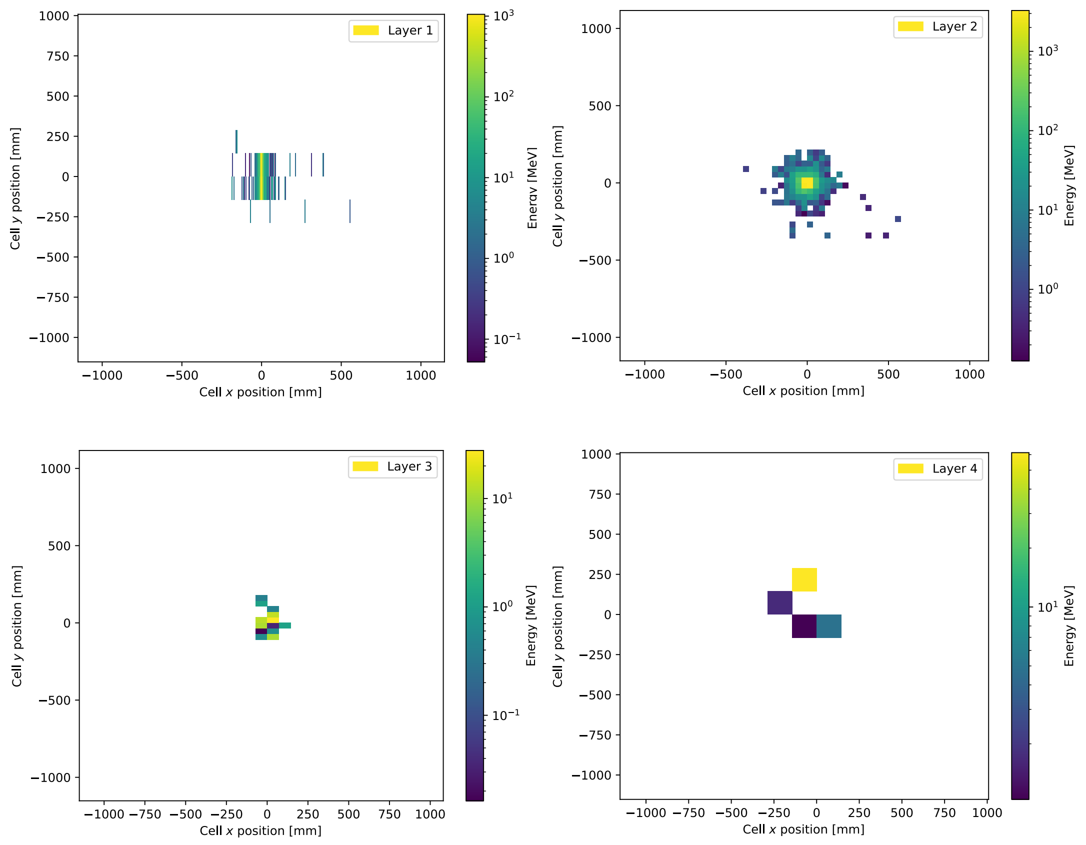
Layer	$dx \times dy$ [mm^2]	$d\eta \times d\phi$	Total X0
ECAL 1	4.5 x 144	0.025/8 x 0.1	6
ECAL2	36 x 36	0.025 x 0.025	16
ECAL3	144 x 144	0.1 x 0.1	3

Layer	$dx \times dy$ [mm^2]	$d\eta \times d\phi$	Total λ_{int}
HCAL 1	144 x 144	0.1 x 0.1	1.5
HCAL2	144 x 144	0.1 x 0.1	4.1
HCAL3	288 x 288	0.2 x 0.2	1.8



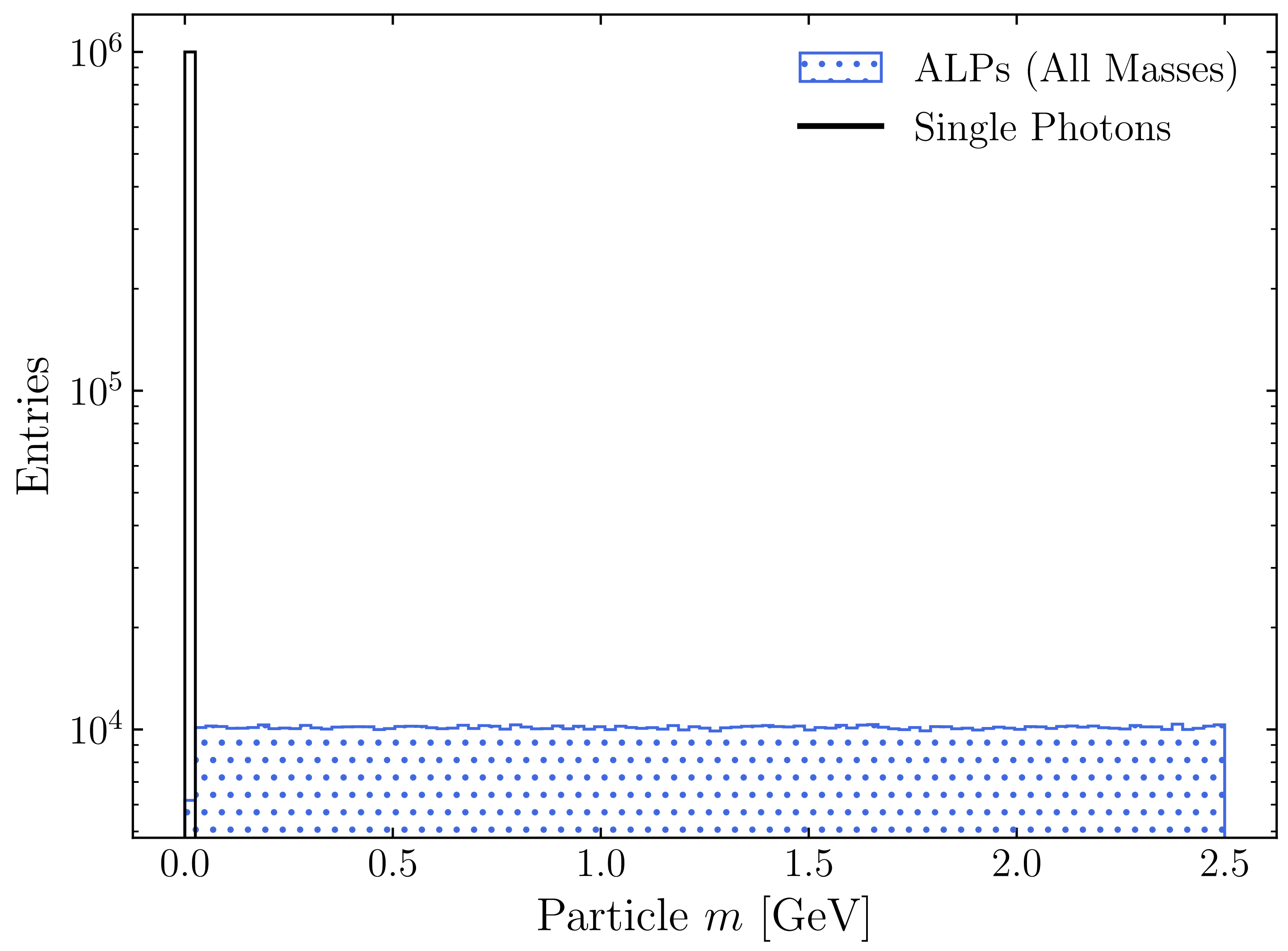


Sample photon cluster

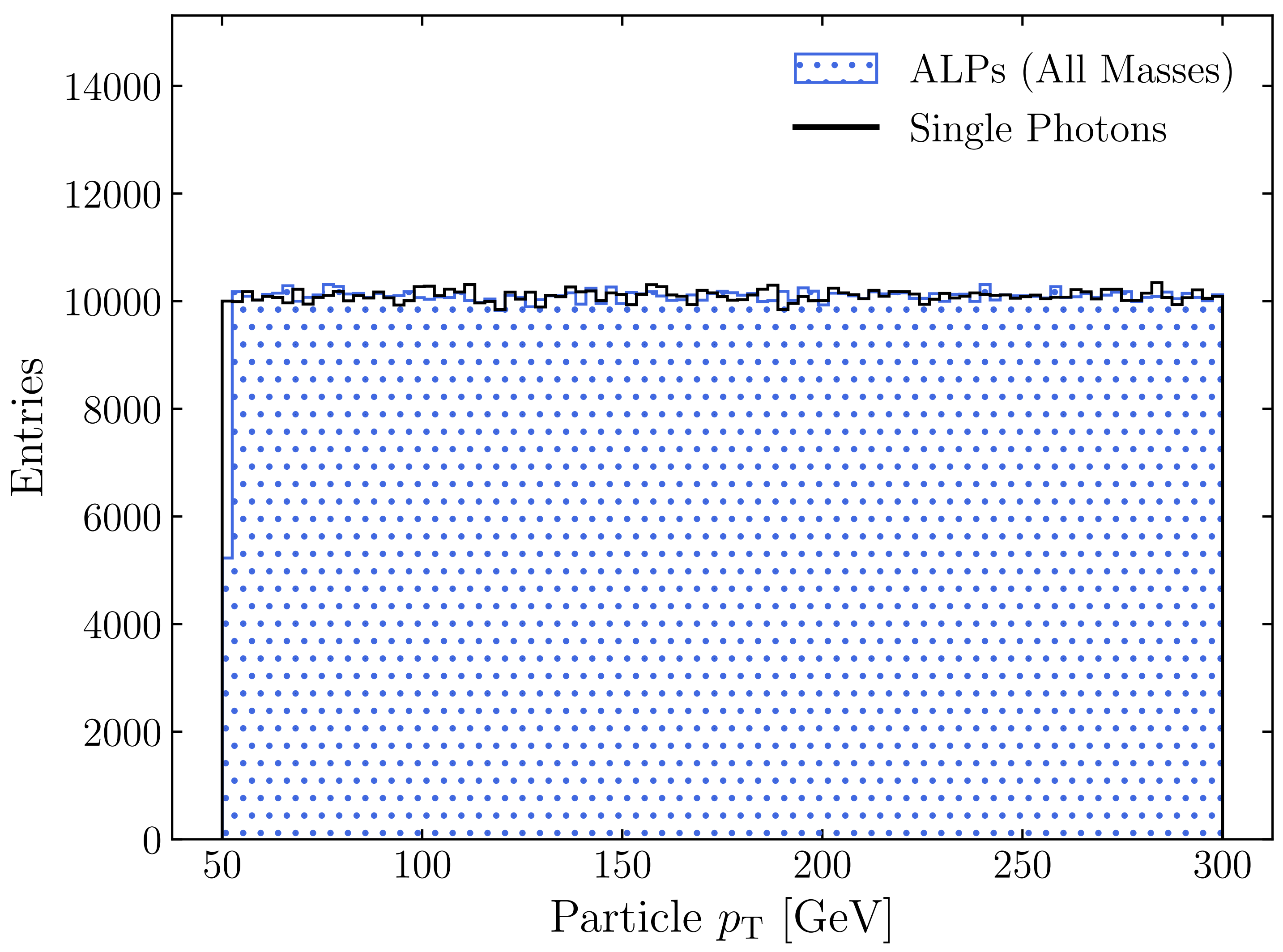


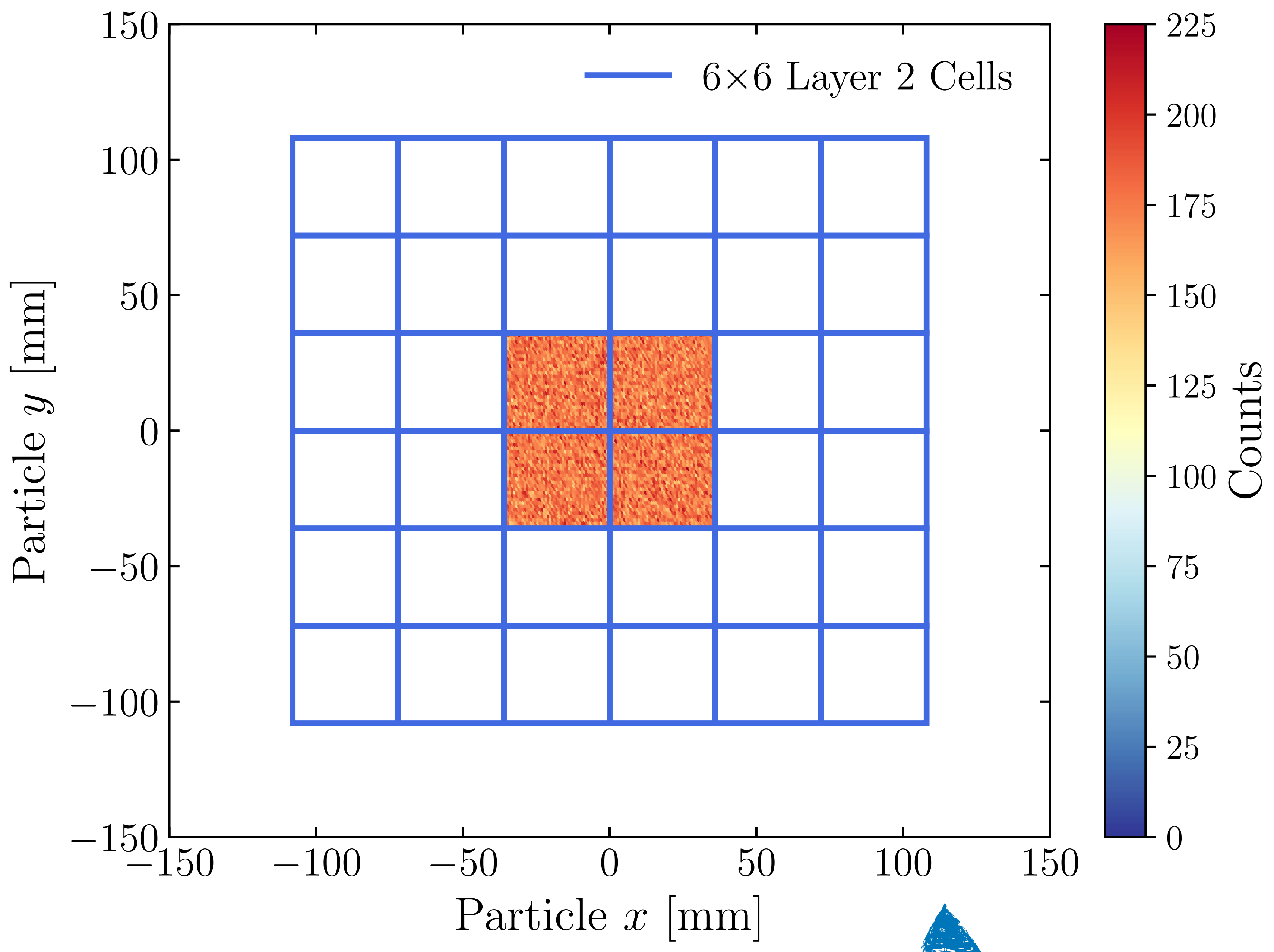
Benchmark Dataset: Details

ALPs generated to cover (0.01, 2.5) GeV masses uniformly



Both ALPs and single photon momenta uniformly distributed in (50, 300) GeV





Particle gun location smeared within a 2 x 2 block of ECAL2 cells to remove biases on particle incidence

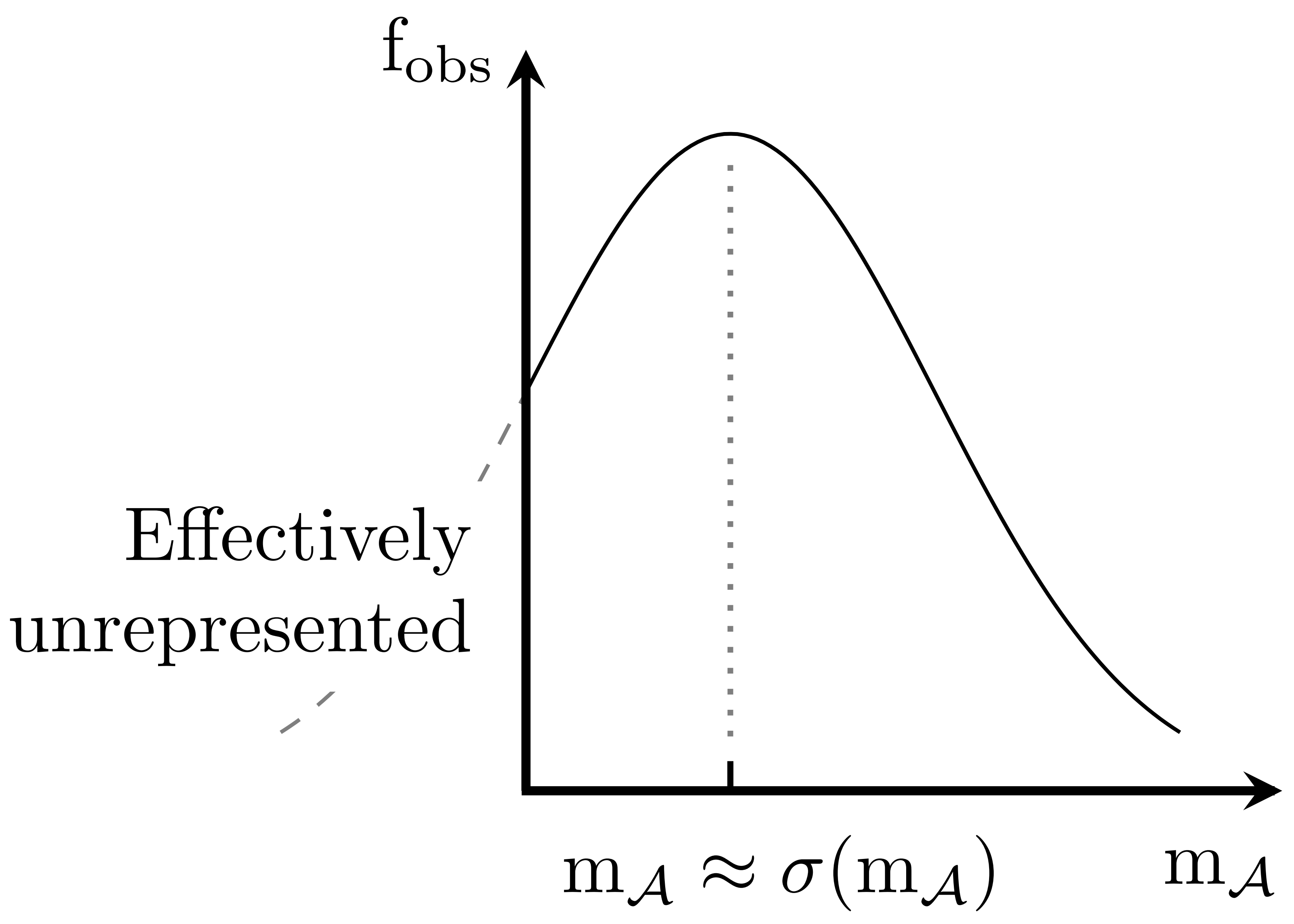
How Does a Transformer Work?

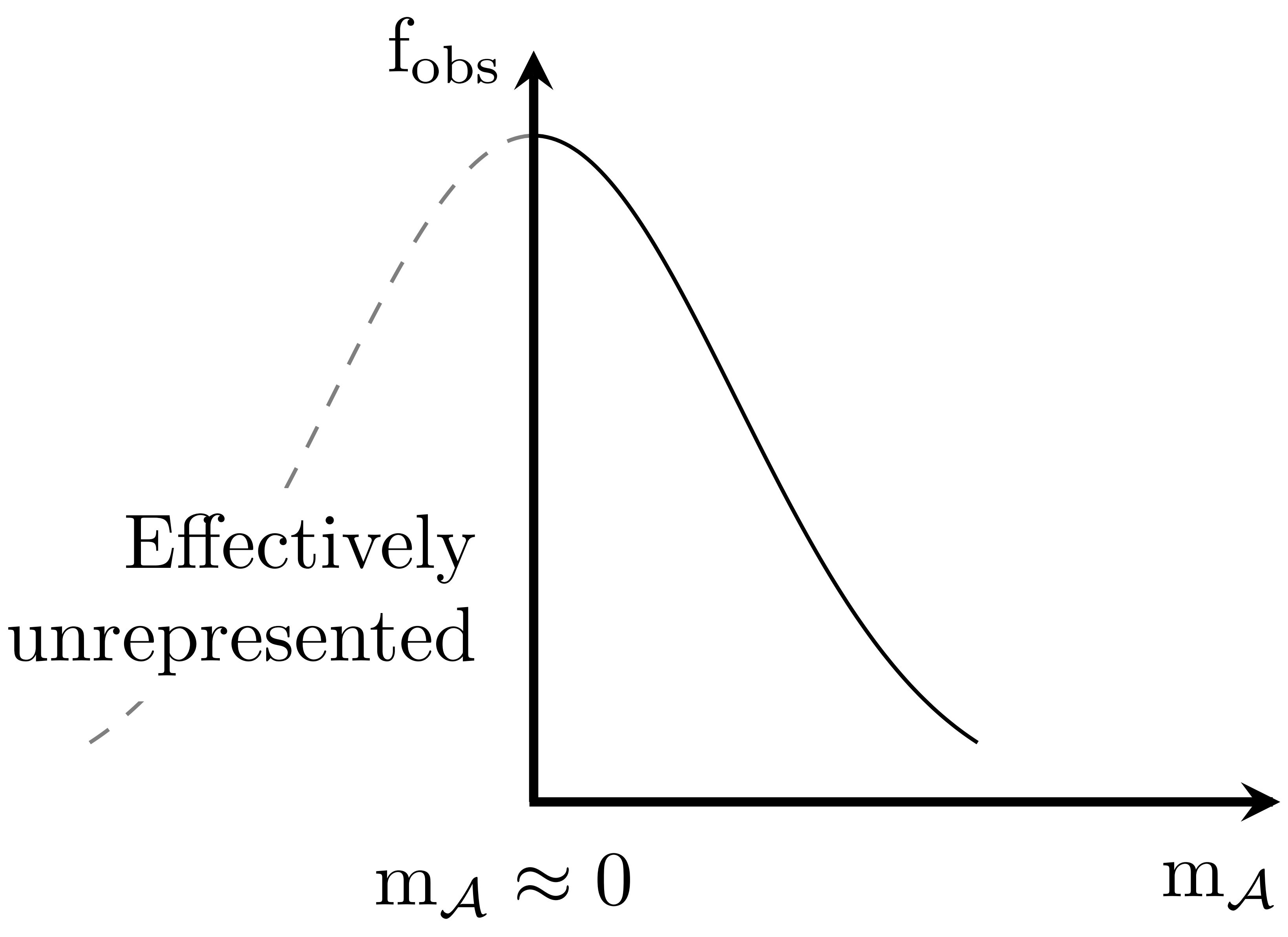
$$\text{Attention} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d^k}}\right)V$$

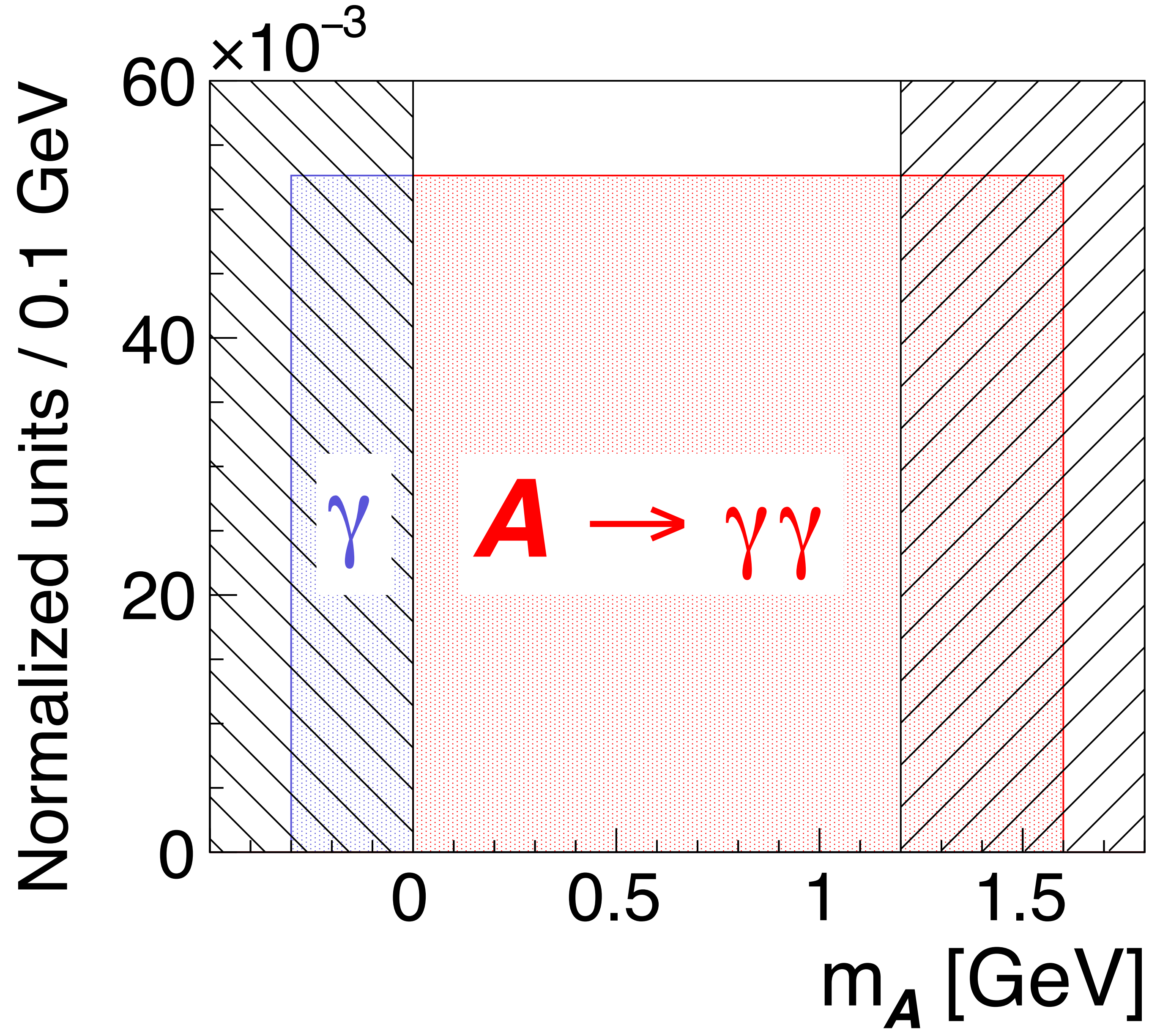
Phys. Rev. D 108 (2023) 052002 [link]

Domain Continuation

- Extend regression targets into **unphysical range** (e.g. giving photons a small negative mass) to prevent biases from **effectively unrepresented** observables due to regressing on $m_a < \sigma(m_a)$
  - Allows for modeling as $m_a \to 0$, reject regressed values with $m_a < 0$







Benchmark Models: Details

	SSV-level		Cell-level	
	BDT	**DNN**	**CNN**	**PFN**
Inputs	13 shower shape variables	13 shower shape variables	Cell (E, η, ϕ) on a 144×64 grid in coordinates after re-binning; EM layers 1–3 treated as different channels, other layers ignored	Point clouds of cell (E, η, ϕ, layer)
Architecture	100 trees, max depth 5; event- and feature-subsampling at 0.8	3 hidden layers × 32 nodes, ReLU; single sigmoid output	3 sparse-conv blocks (5×5, then 3×3 kernels), 32 → 128 channels, each with BatchNorm + ReLU + 2×2 max-pool; dense head	$F(\sum_i \Phi(p_i))$: Φ and F each 6 fully-connected layers of width 128, ReLU; global sum pooling
Training/Loss	Binary logistic objective	Batch size of 5000; trained for 50 epochs with binary cross-entropy	Batch size of 16; trained for 5 epochs with multi-task L = 0.8·BCE + 0.2·MSE (mass regression, signal only)	Batch size of 256; trained for 50 epochs with binary cross-entropy
AUC	**0.92**	**0.92**	**0.95**	**0.97**