Cross-Domain Transfer with Particle Physics Foundation Models: From Jets to Neutrino Interactions

Gregor Krzmanc (Stanford/SLAC),
Vinicius Mikuni, Benjamin Nachman, Callum Wilkinson

https://arxiv.org/abs/2604.12364

Stanford University

Foundation models in particle physics

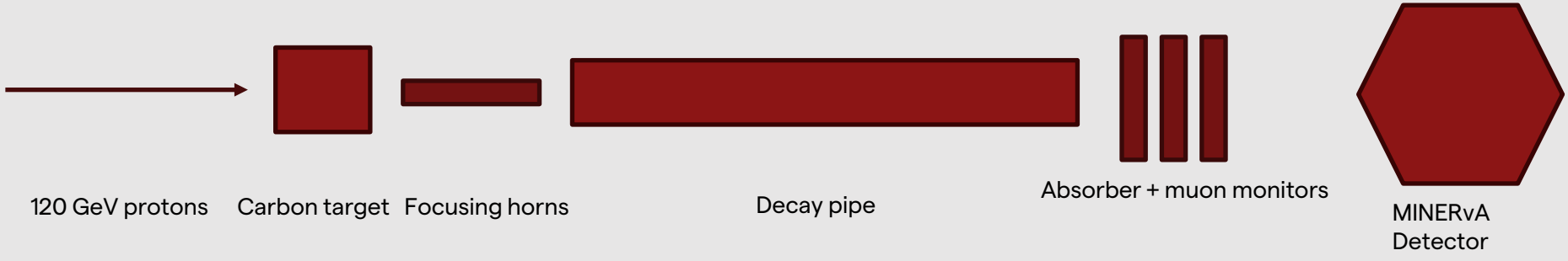
- Foundation models – models pretrained on large amounts of data, usable for a wide range of downstream tasks
- Existing efforts at the LHC experiments: OmniLearned [2510.24066], Masked Particle Modeling [2401.13537]...
- The purpose of this work**: Can models pre-trained on collider jets transfer knowledge to a medium-energy fixed-target neutrino experiment?**

Outline of this talk

1. Foundation models in particle physics
2. The MINERvA experiment at Fermilab
3. Defining tasks and baselines on MINERvA Open Data
4. Applying ML models to MINERvA Open Data

The MINERvA experiment at Fermilab

(Main Injector Neutrino ExpeRiment to study ν-A interactions)



Focus produced mesons (π^+, π^-)

$\pi^+ \rightarrow \mu^+ + \nu_\mu$ (Forward-horn current)

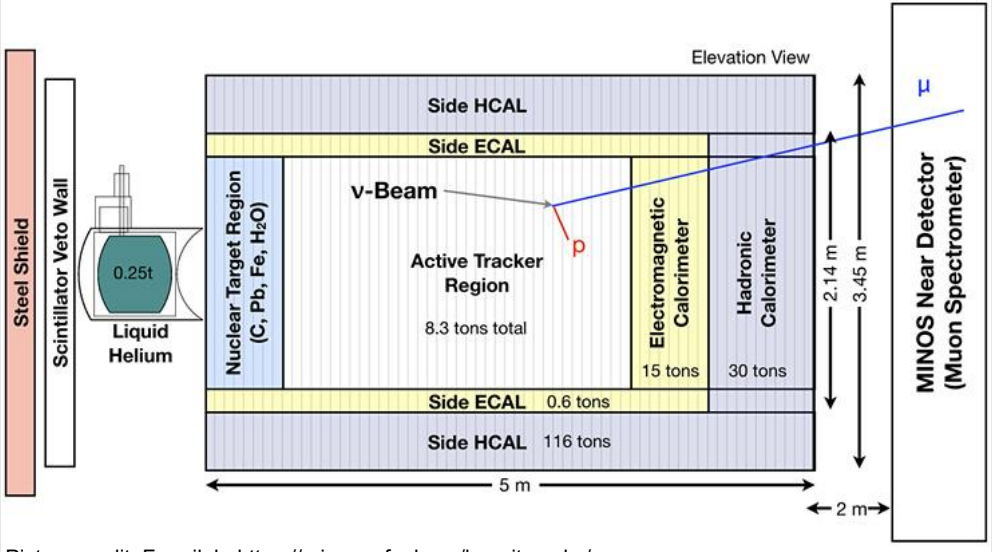
$\pi^- \rightarrow \mu^- + \overline{\nu_\mu}$ (Reverse-horn current)

675 m

210 m

MINERvA Detector

- Precise measurements of neutrino-nucleus interactions
- Fileable liquid helium and water tanks



Picture credit: Fermilab, https://minerva.fnal.gov/how-it-works/

Charged current (CC) events:
Muon in the final state
$\nu_\mu + Y \rightarrow \mu^- + X$

Y = neutron (for quasi-elastic), nucleus…
X = hadronic system, proton (for quasi-elastic)

Neutral current (NC) events:
$\nu_\mu + Y \rightarrow \nu_\mu + X$

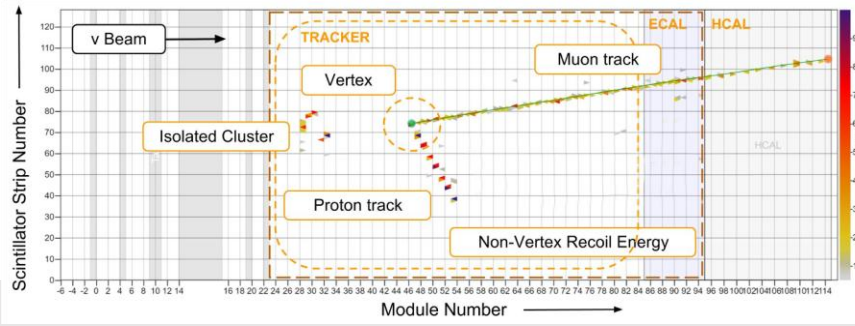
MINERvA Open Data [https://www.osti.gov/biblio/3022562]

- Contains both MC and real data
- FHC (forward horn current); RHC (reverse horn current)
- Medium-energy beam (avg. neutrino energy ~ 6 GeV); lower energies will be provided soon too
- Contains both real data and simulation
- FHC MC: ~11 TB of MC data, ~4000 columns containing info about final-state particles, various truth-level links, reconstructed objects, and analysis variables

Medium Energy Neutrino (FHC) Playlists

Playlist Name	Water Target status	Helium Target status	Data file list	Monte Carlo file list	Data POT ($\times 10^{20}$)
minervame1A	Empty	Empty	Data 1A	MC 1A	0.90
minervame1B	Empty	Full*	Data 1B	MC 1B	0.19
minervame1C	Empty	Full	Data 1C	MC 1C	0.43
minervame1D	Empty	Full	Data 1D	MC 1D	1.4
minervame1E	Empty	Full	Data 1E	MC 1E	1.0
minervame1F	Empty	Full	Data 1F	MC 1F	1.7
minervame1G	Empty	Empty	Data 1G	MC 1G	1.4
minervame1L	Full	Empty	Data 1L	MC 1L	0.13
minervame1M	Full	Empty	Data 1M	MC 1M	2.1
minervame1N	Full	Empty	Data 1N	MC 1N	1.1
minervame1O	Full	Empty*	Data 1O	MC 1O	0.30
minervame1P	Full	Full	Data 1P	MC 1P	0.47

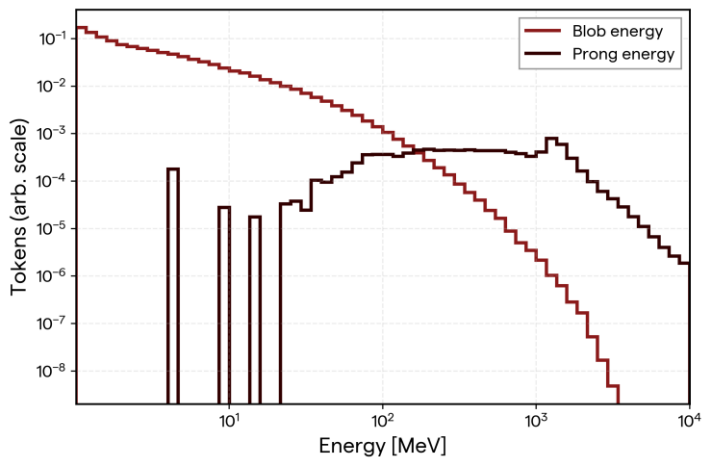
MINERvA Open Data



We represent events as a **set of reconstructed objects**:

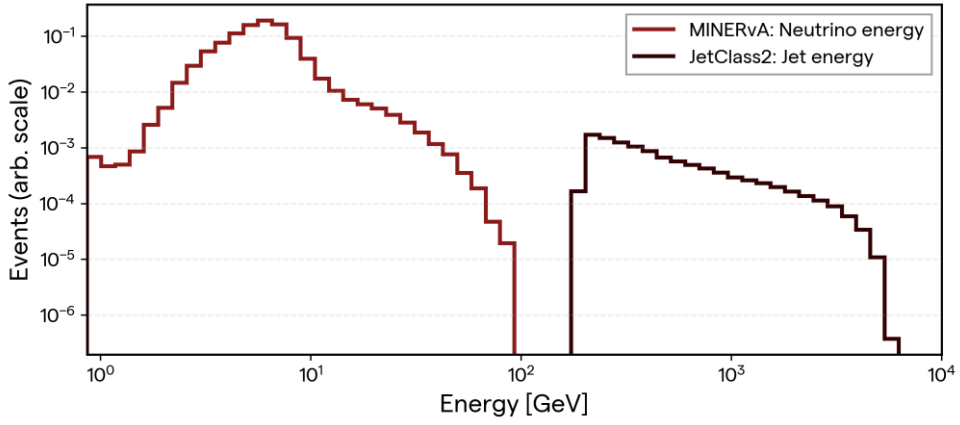
- **Prongs:** "tracks" with position and momentum + basic PID (proton/pion/muon/unknown)
- **Blobs:** energy deposits without directionality and no PID hypothesis
- **Photons:** up to two reconstructed photons
- **Muons:** 0 or 1 reconstructed muon

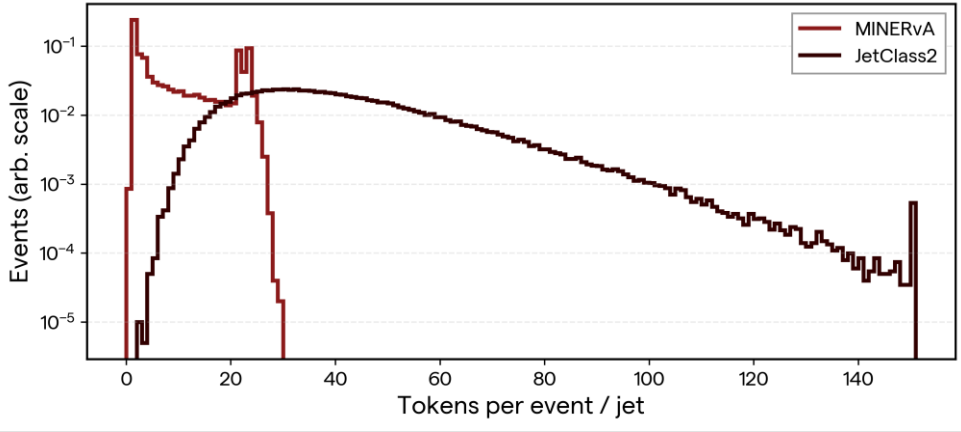
+ global event-level features



Collider jets vs. neutrino events

- OmniLearned: pretrained on 1B jets
- Mismatch in energy scale and event multiplicity





Tasks on MINERvA events

Classification

- Using 3 separate signal definitions:

$\mathrm{CC1}\pi^{\pm}$ $\mathrm{CC1}\pi^{0}$ $\mathrm{CCN}\pi^{\pm}$

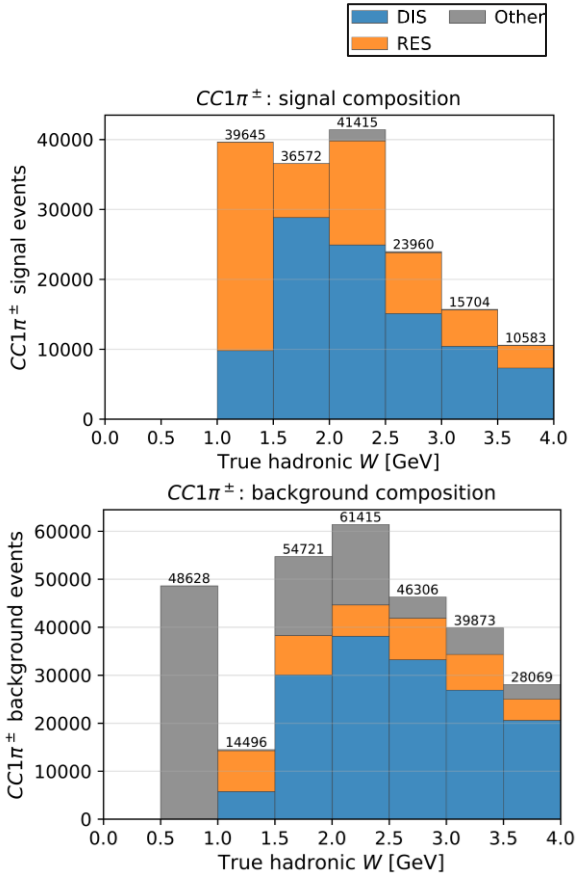
- Backgrounds:

NC (no muon in final state)

CC (all classes except signal class)

- Report scores in bins of true W

(invariant mass of the hadronic system) and energy of the charged pion

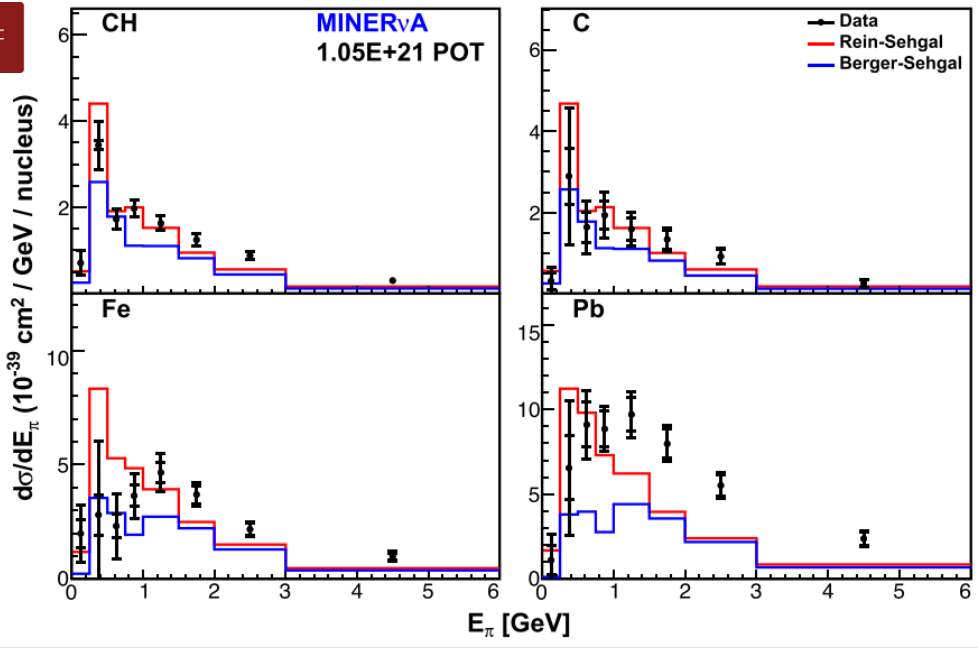


Tasks on MINERvA events

Classification

- Comparing different nuclear interaction models *[Phys. Rev. Lett. **131**, 051801, 2023]*

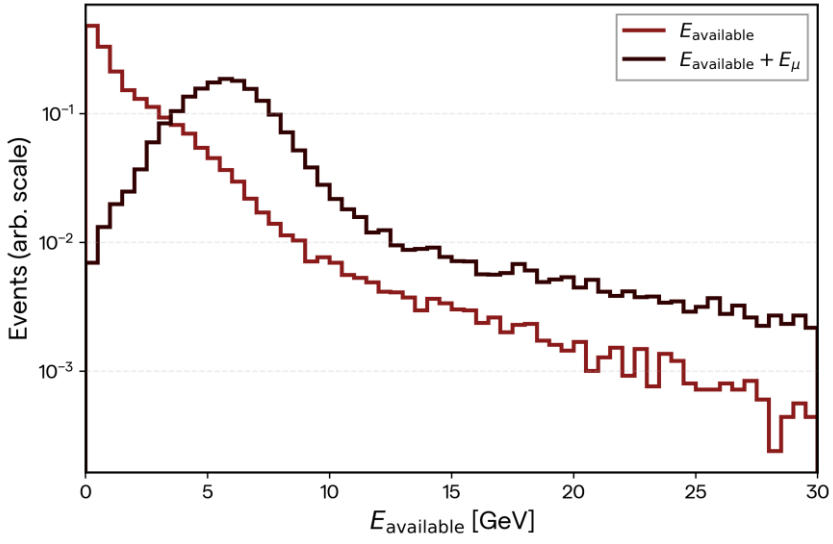
$CC1\pi^{\pm}$

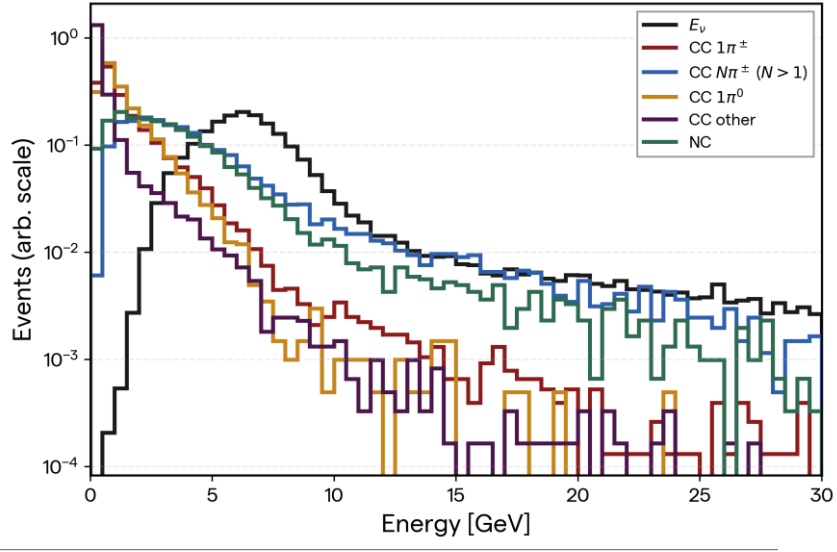


Tasks on MINERvA events

Regression

- Regress available energy $E_{available}$ - sum of energies of protons, pions, electrons, photons, charged kaons – excludes poorly modeled neutrons, strange baryons etc.
- Such definition is less model-dependent than E

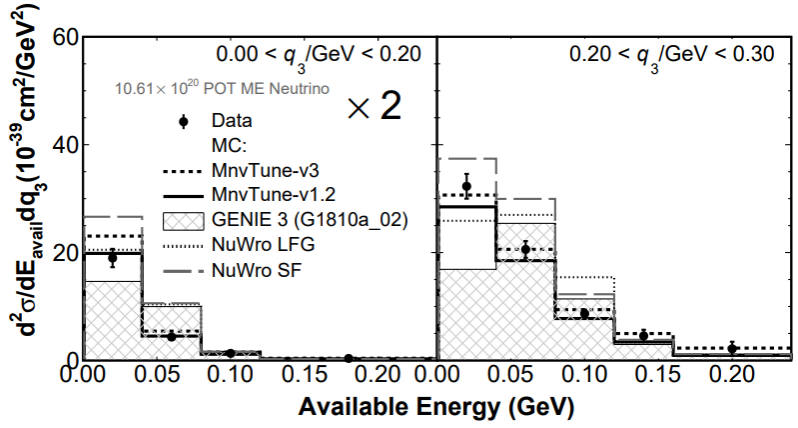




Tasks on MINERvA events

Regression

- Charged current cross-section measurements: $\frac{d^2\sigma}{dE_{available}dq_3}$ (q_3: three-momentum transfer) at low q_3 (< 1.2 GeV) *[Ascensio et al., 2022, https://inspirehep.net/literature/1952230]*
- At low q_3, nuclear effects dominate – precise cross-section measurements help constrain nuclear models



Baselines

- Available energy regression: sum of energies from EM components of the detector * correction factor (=1.17) – from MINERvA code
- Classification – cut-based baseline

$CC1\pi^{\pm}$ 1 charged prong, 1 muon, 1+ Michel tag

$CC1\pi^{0}$ 1 muon, 2 photons with $m_{\gamma\gamma} < 2m_{\pi 0}$

$CCN\pi^{\pm}$ 1 charged prong, 1 muon, 1+ Michel tag

ML models

- **Models using event-level + particle-level features:**
  - **OmniLearned**: Point Edge Transformer trained on ~1B jets using classification and generation tasks [2510.24066]
  - **HyperScale**: ViT-like transformer trained on ~1B jets using classification. Using positional encodings with η, φ coordinates [2606.19781]
  - Input = Particle tokens + "Event" Token (global features)
  - Read out the event-level labels from a CLS token
- **MLP**: operating only on event-level features

Model inputs

- During pre-training (done by the authors of OmniLearned and HyperScale):

Kinematic features
$\log p_T, \log E, \Delta\eta, \Delta\phi$

PID
(learned lookup table)

Additional features
(track/vertex information…)

Pre-training tasks

OmniLearned

- Classification (jet tagging)
- Jet generation

HyperScale

- Classification (jet tagging)

Model inputs

- During pre-training (done by the authors of the models OmniLearned and HyperScale):

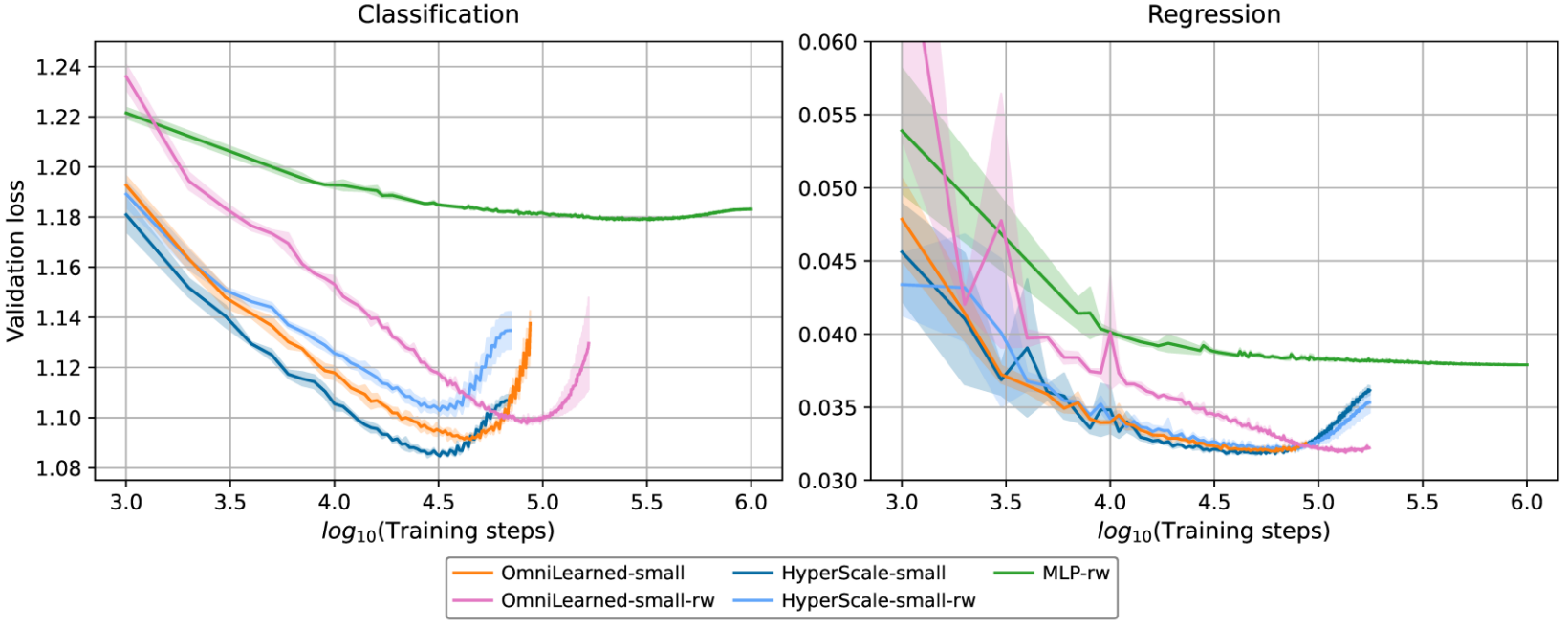
Kinematic features	PID	Additional features
$\log p_T, \log E, \Delta\eta, \Delta\phi$	Learned lookup table	(track/vertex information…)

- Fine-tuning on MINERvA events:

Kinematic features	PID	Additional features
$\log p_T, \log E, \Delta\eta, \Delta\phi$	(prong/blob/γ/μ) Learned lookup table	$\langle dE/dx \rangle, x, y, z, t$

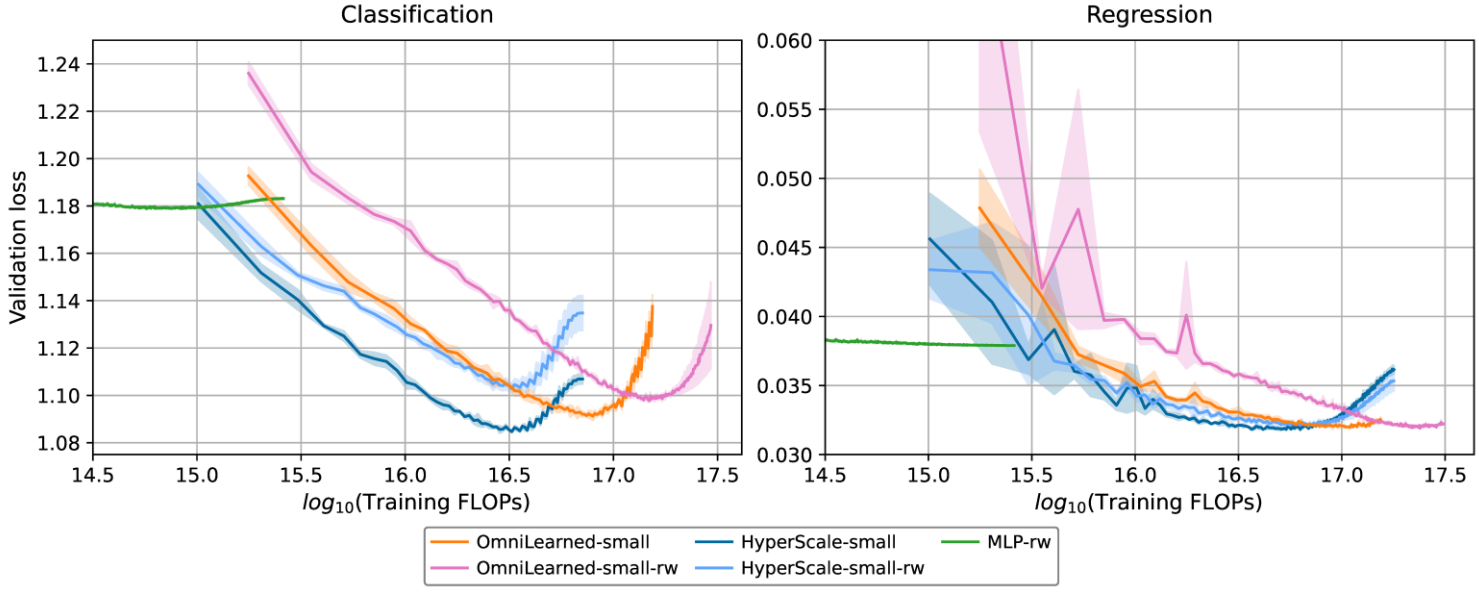
Training dynamics

- Pre-trained models (OmniLearned / HyperScale) learn with less steps than scratch-trained models with the same architecture (-rw)
- “Small” scale (~3M parameters for each model)



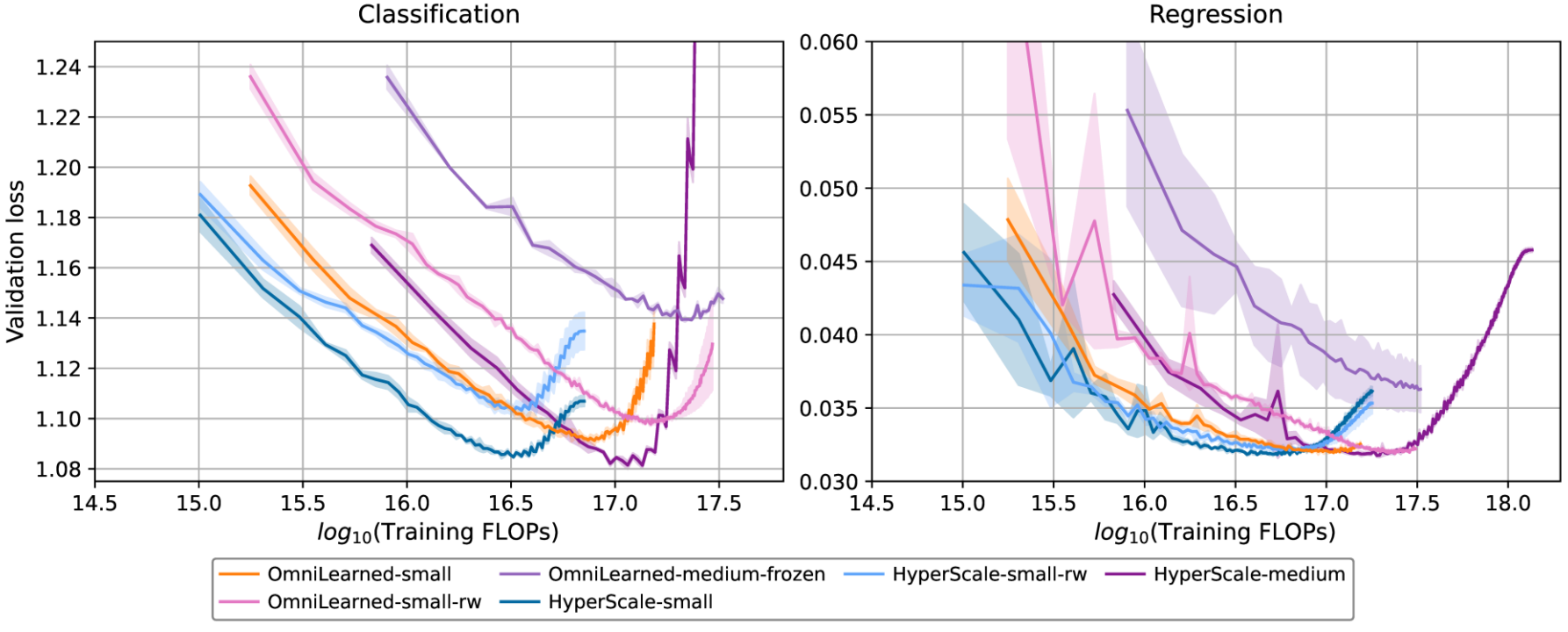
More interesting: Compute vs. performance

- Same amount of compute leads to better performance using a pre-trained model compared to a scratch-trained model with the same architecture (rw = random weights)



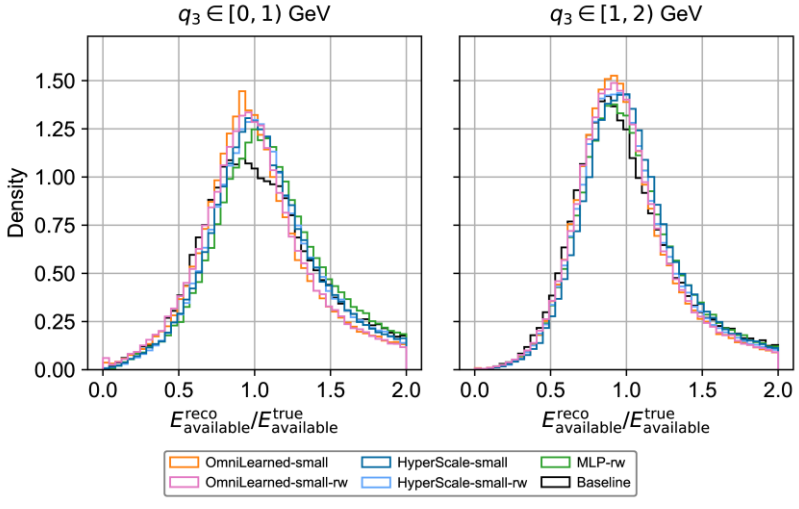
Scaling

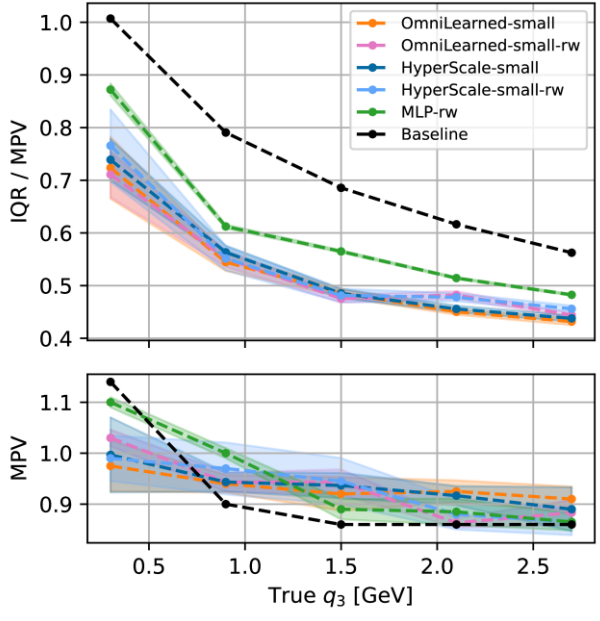
- Larger model can lead to better performance (HyperScale-medium (15M) vs. HyperScale-small (3M))
- OmniLearned-medium-frozen: frozen backbone, fine-tune only the heads



Results - Regression

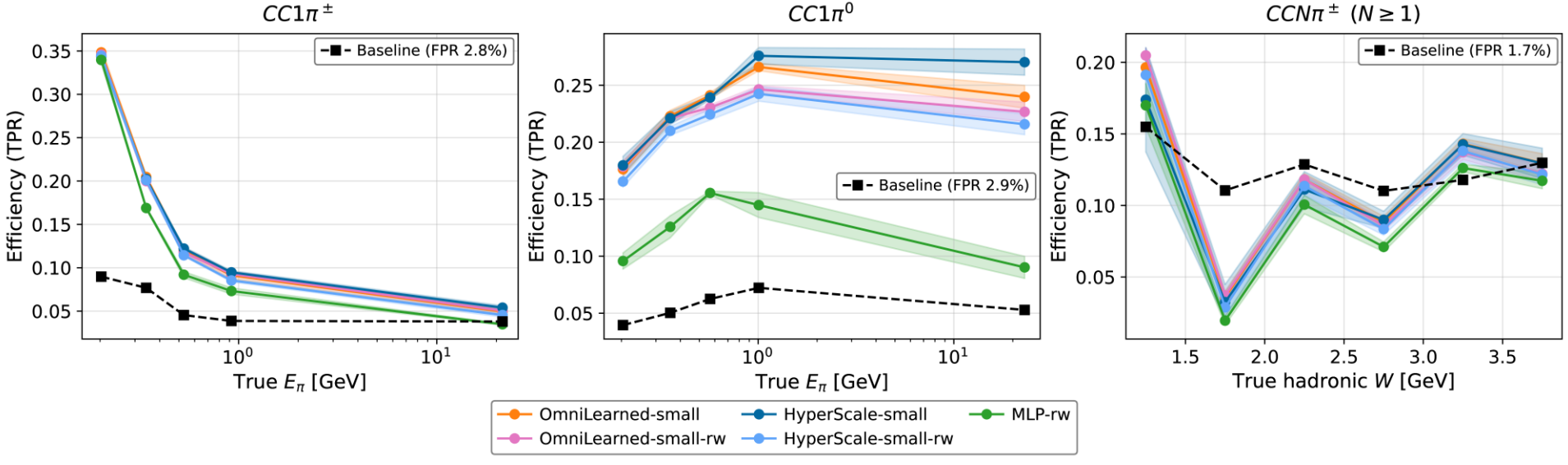
- Report IQR of the predicted E / true E distribution in bins of three-momentum transfer q_3
- Event selection similar to MINERvA's CC-inclusive analysis (clean muon + some additional criteria)





Results - Classification

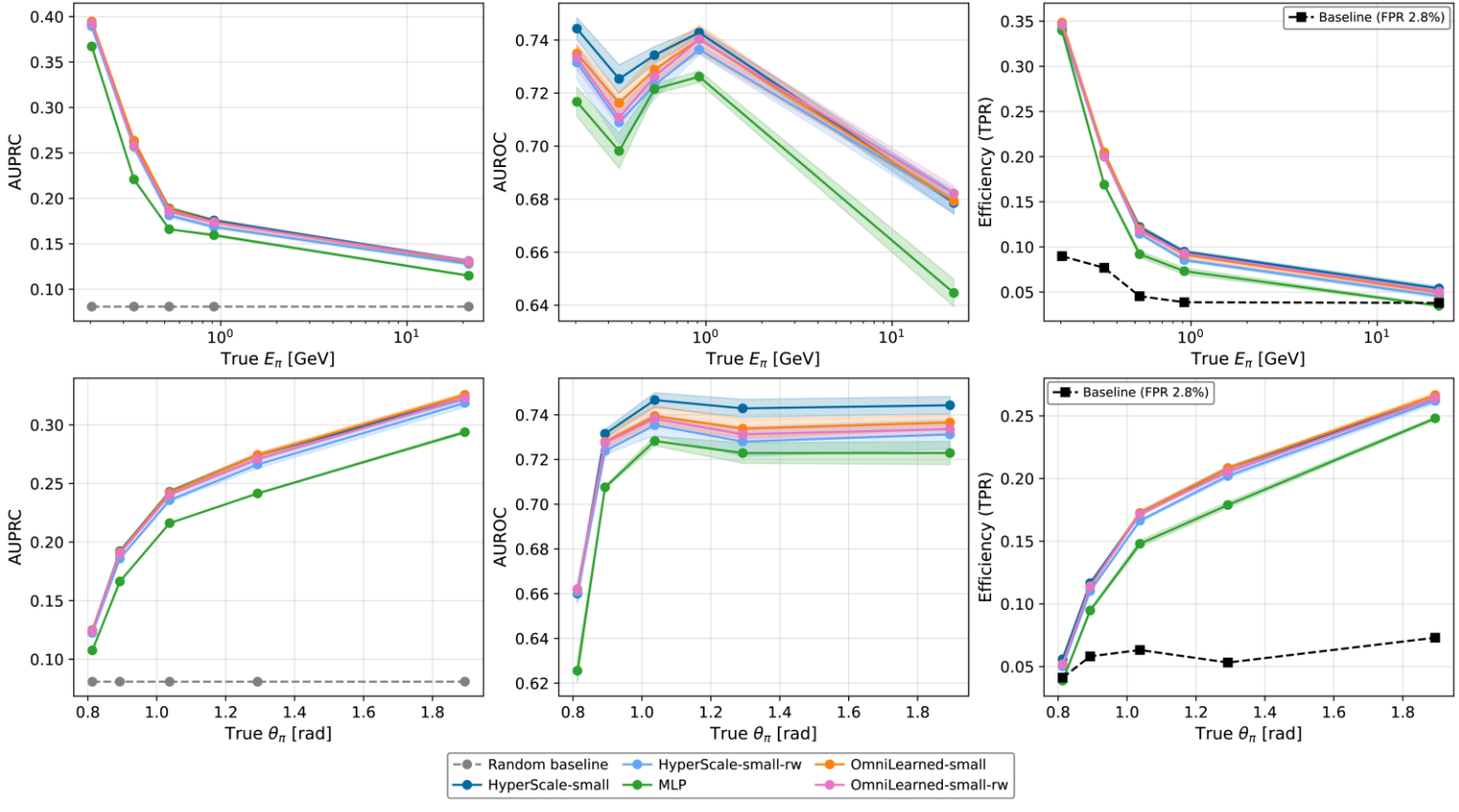
- Report TPR at the class threshold giving the FPR of the baseline
- Event-level features include everything used by the cut-based baseline - the performance of the models
- We do “global” loss weighting, might be better to do per-bin weighing or so – in some bins for the N pion case, the model always predicts background and no signal



Results - Classification

- In terms of pion angle and energy

$CC1\pi^{\pm}$



What if we use a pretrained language model?

- Large Language Models — the Future of Fundamental Physics? [2506.14757]
  - Using SKA data (3D grid) to predict cosmological parameters
  - Using Qwen2.5-0.5B weights leads to faster training of the models
- Maybe our random initialization is bad and pre-training on any data is good?

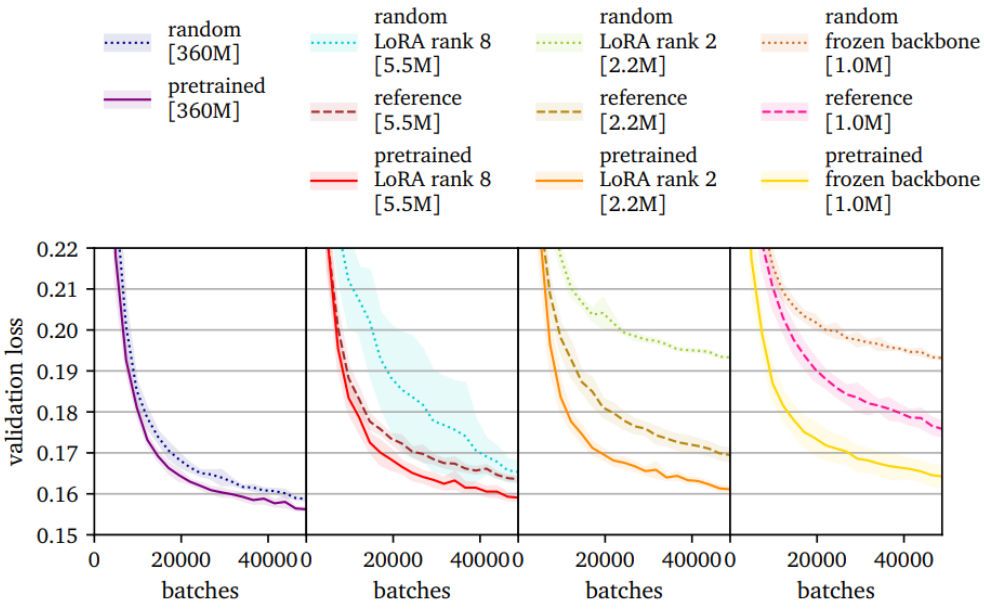
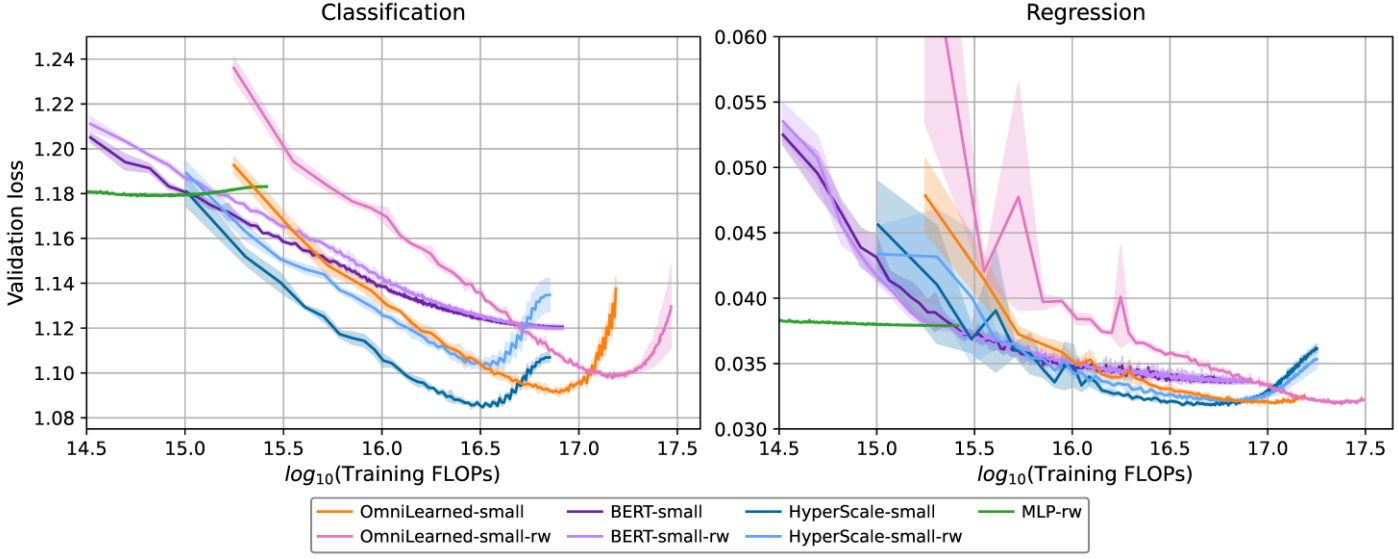


Figure 8: Mean validation loss with a 1σ-band determined from 5 runs.

Results using BERT

- Using small BERT with 3M parameters (prajjwal1/bert-tiny)
- No tokenization - project feature vectors with a learnable linear projection in the transformer embedding space
- Gap pretrained vs. random init. (rw) much smaller compared to OmniLearned and HyperScale
- Our results are with a much smaller model then the SKA paper (3M vs 500M parameters)



Conclusions

- **MINERvA-specific:** Using a reconstructed object-level picture of events seems to help with regression and classification tasks (MLP using only event-level features vs. any transformer model using event-level + particle-level features)
- **More general:** Models pre-trained on jets can acquire inductive biases that generalize across detectors and energy regimes

Details about the dataset

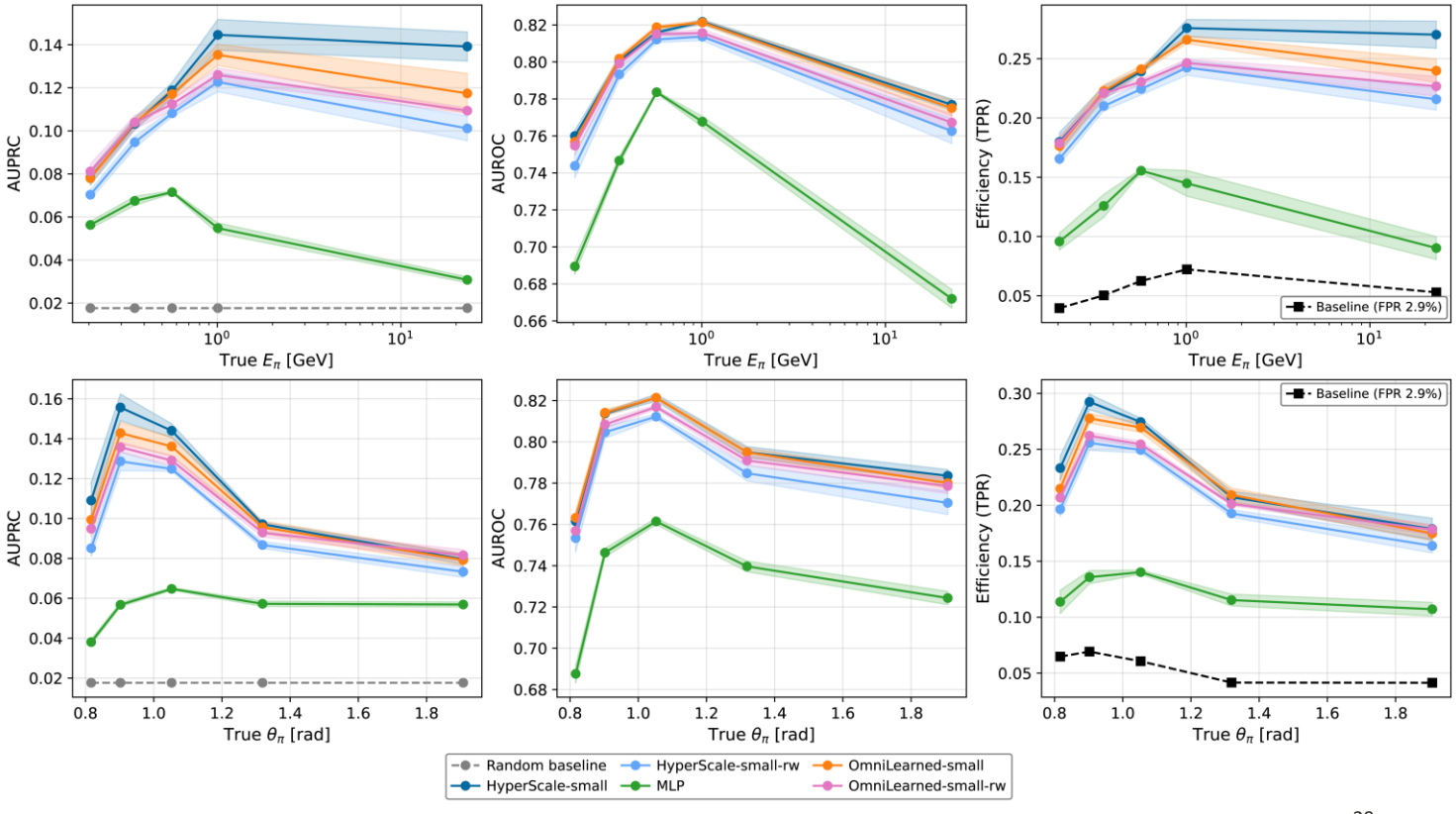
- Playlist 1B (He target full): 2M test, Playlist 1A (both He and H_2O targets empty): 6M training events, 700k validation, 700k events

- Event-level features: muon fuzz energy, muon isolated blobs energy, Minerva's estimator of energy transferred to the hadronic system (E_recoil), number of tagged Michel electrons, log of summed energies for each node type (prongs, blobs...)

Results - Classification

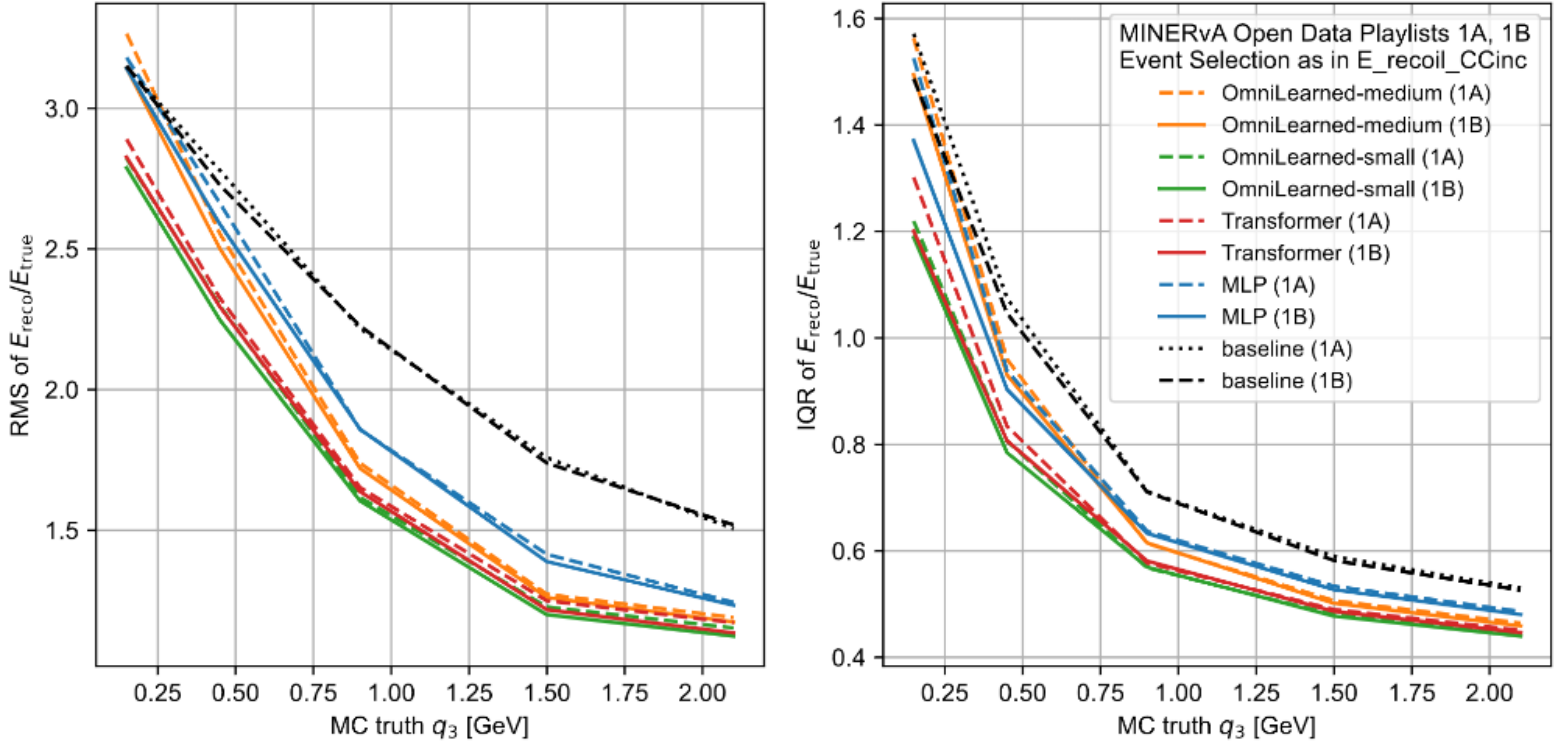
- In terms of pion angle and energy

$\mathrm{CC1}\pi^0$

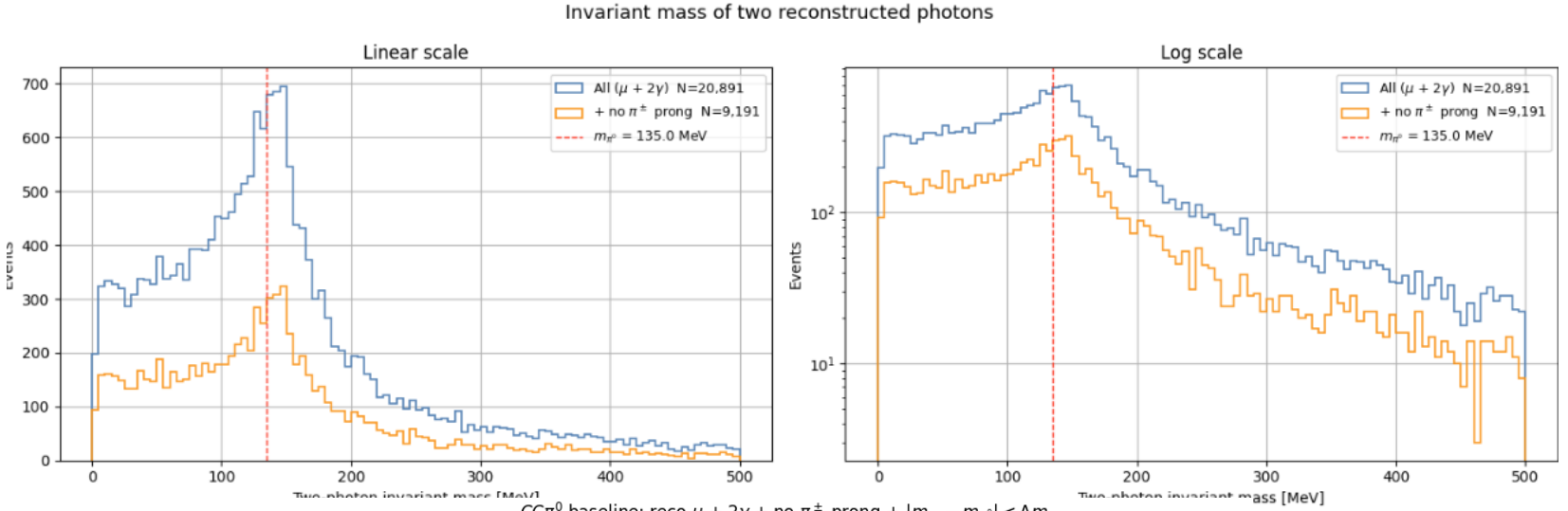


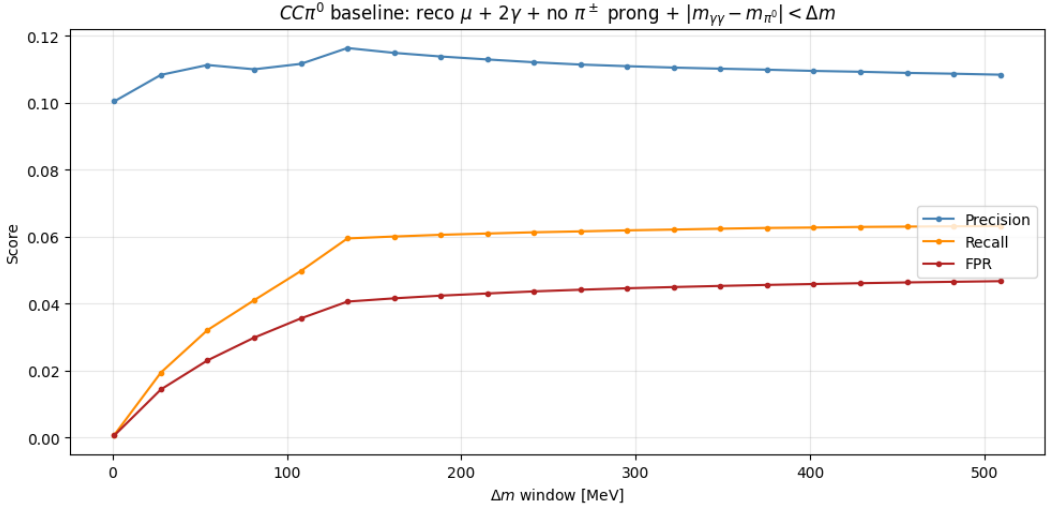
Available Energy Regression

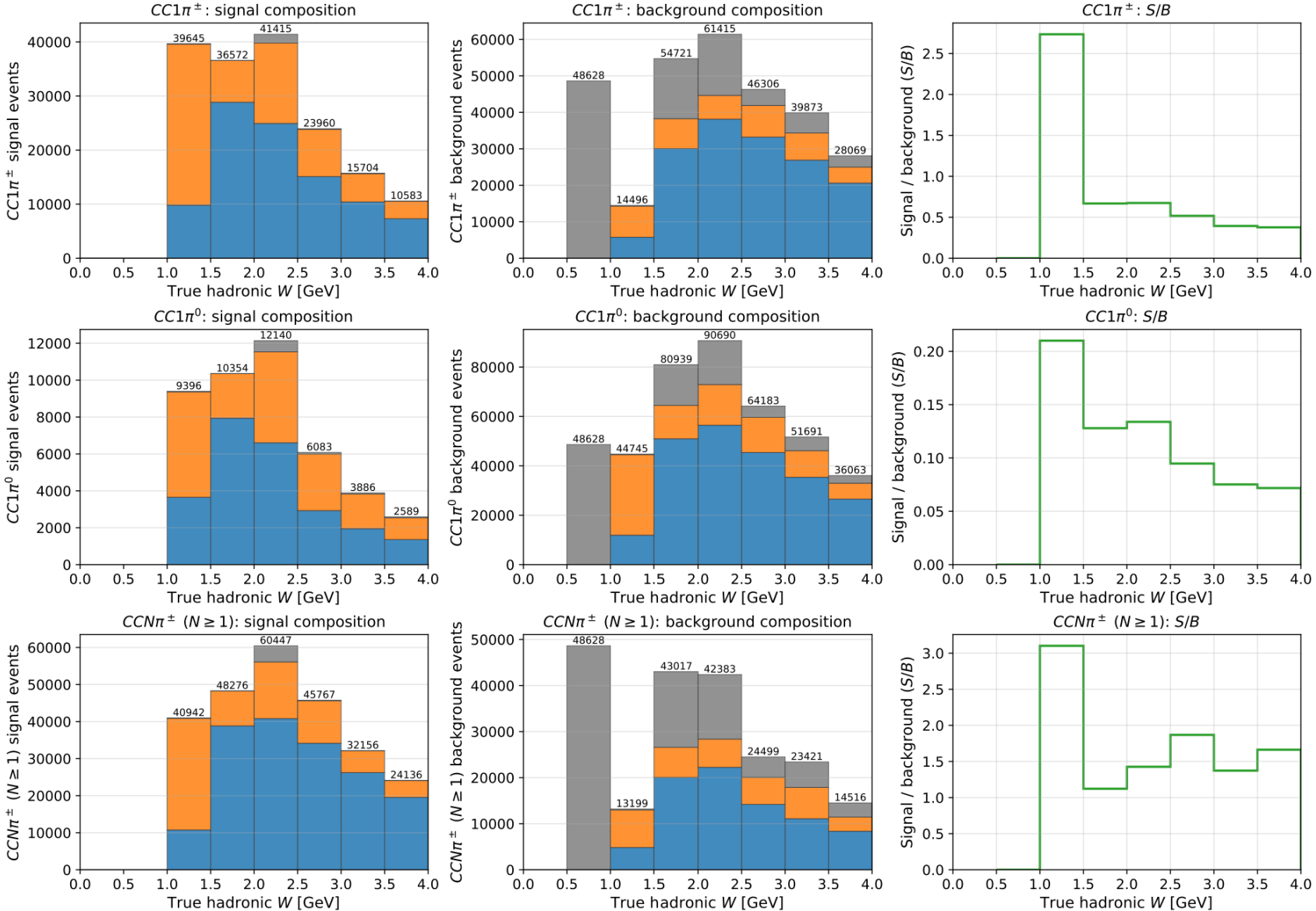
- 1B and 1A playlists –very similar performance



CCPi0 tagging baseline






Event composition vs. true W - MINERvA Open Data Playlist 1A
CC1π±: signal composition
CC1π±: background composition
CC1π±: S/B
CC1π0: signal composition
CC1π0: background composition
CC1π0: S/B
CCNπ± (N ≥ 1): signal composition
CCNπ± (N ≥ 1): background composition
CCNπ± (N ≥ 1): S/B
CC1π± signal events
CC1π± background events
CC1π0 signal events
CC1π0 background events
CCNπ± (N ≥ 1) signal events
CCNπ± (N ≥ 1) background events
Signal / background (S/B)
True hadronic W [GeV]
39645
36572
41415
23960
15704
10583
48628
14496
54721
61415
46306
39873
28069
9396
10354
12140
6083
3886
2589
48628
44745
80939
90690
64183
51691
36063
40942
48276
60447
45767
32156
24136
48628
13199
43017
42383
24499
23421
14516