Large-Scale Pretraining and Finetuning for Efficient Jet Classification

Zihan Zhao, Farouk Mokhtar, Raghav Kansal, Billy Li, Javier Duarte

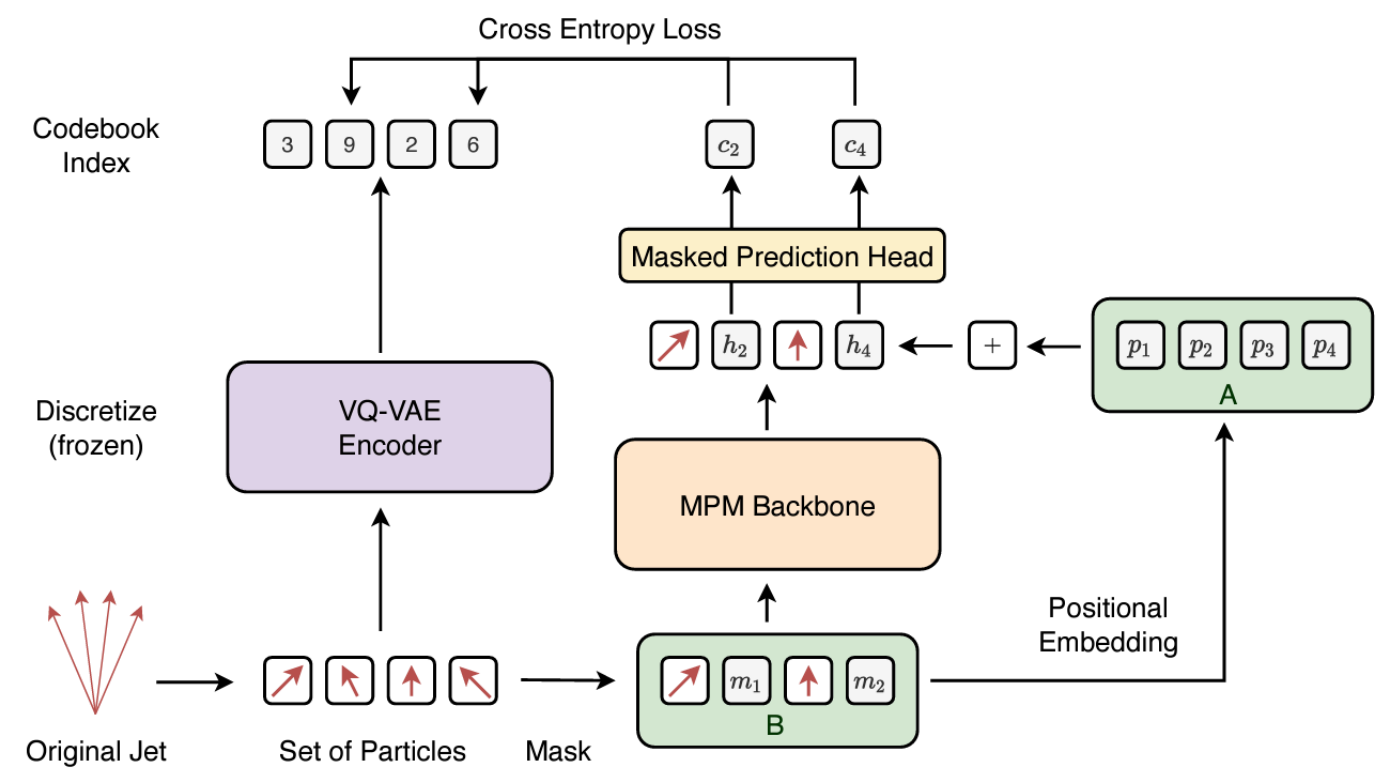
ML4Jets 2024
Nov 6

https://arxiv.org/abs/2408.09343

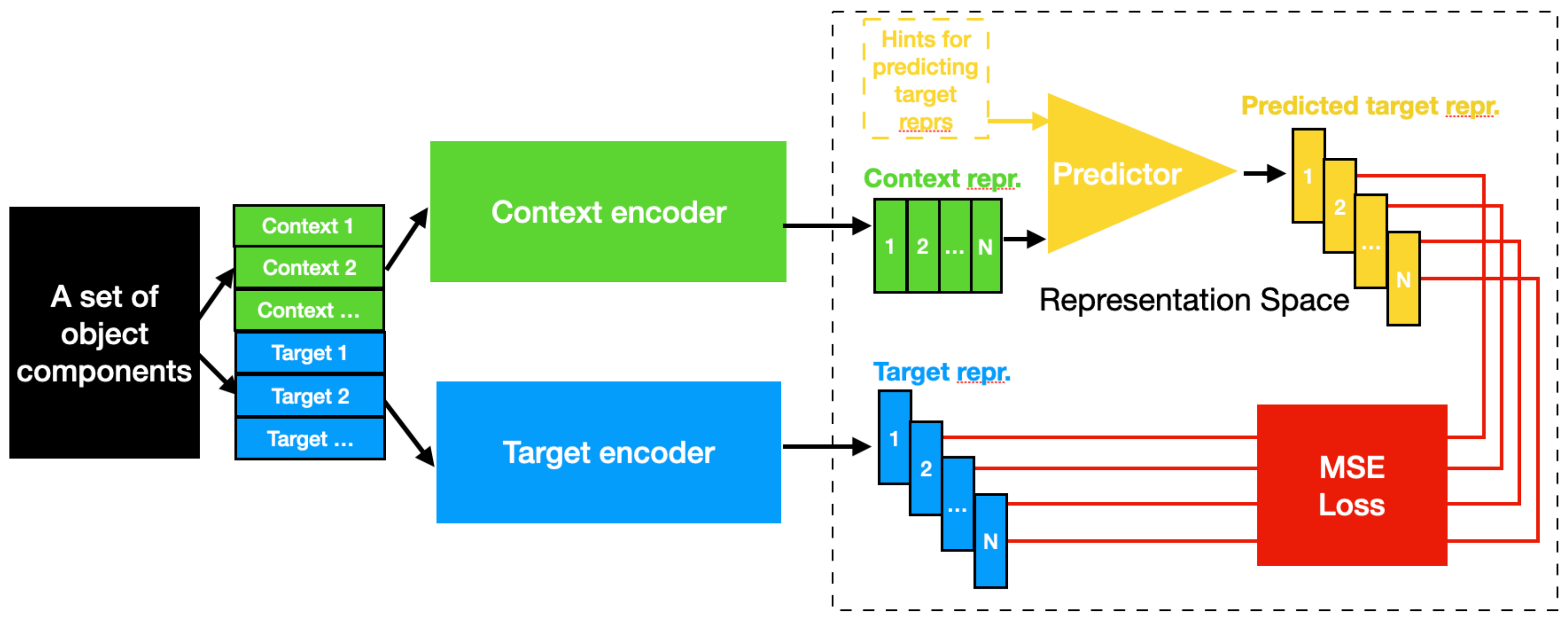
Intro to SSL strategies

As opposed to supervised learning, which is limited by the availability of labeled data, self-supervised approaches can learn from vast unlabeled data (2304.12210)

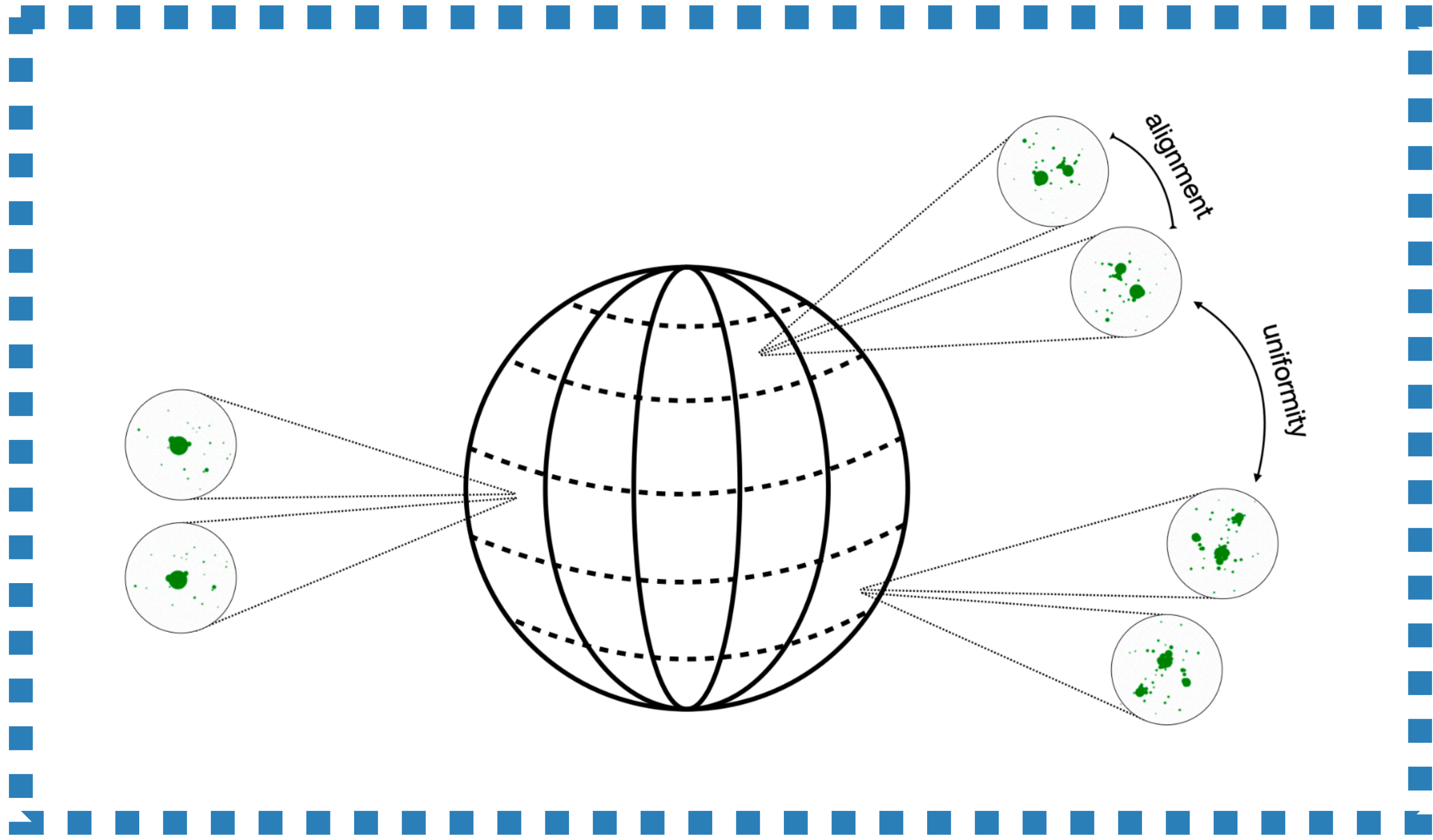
To learn useful features from the data itself without using labels



Masked Modeling



JEPA

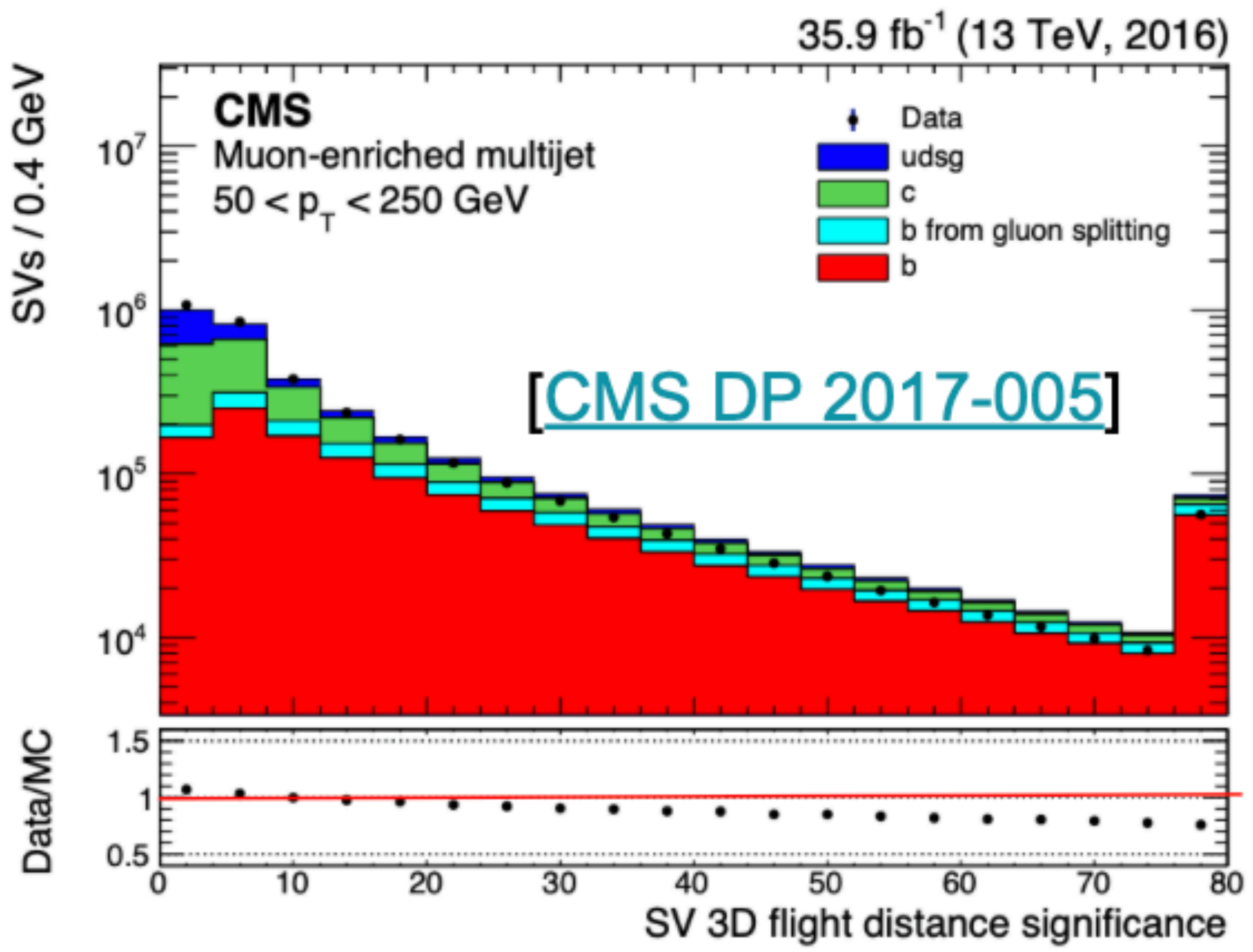


Contrastive Learning

2401.13537 https://indico.cern.ch/event/1386125/contributions/6083379/ 2108.04253

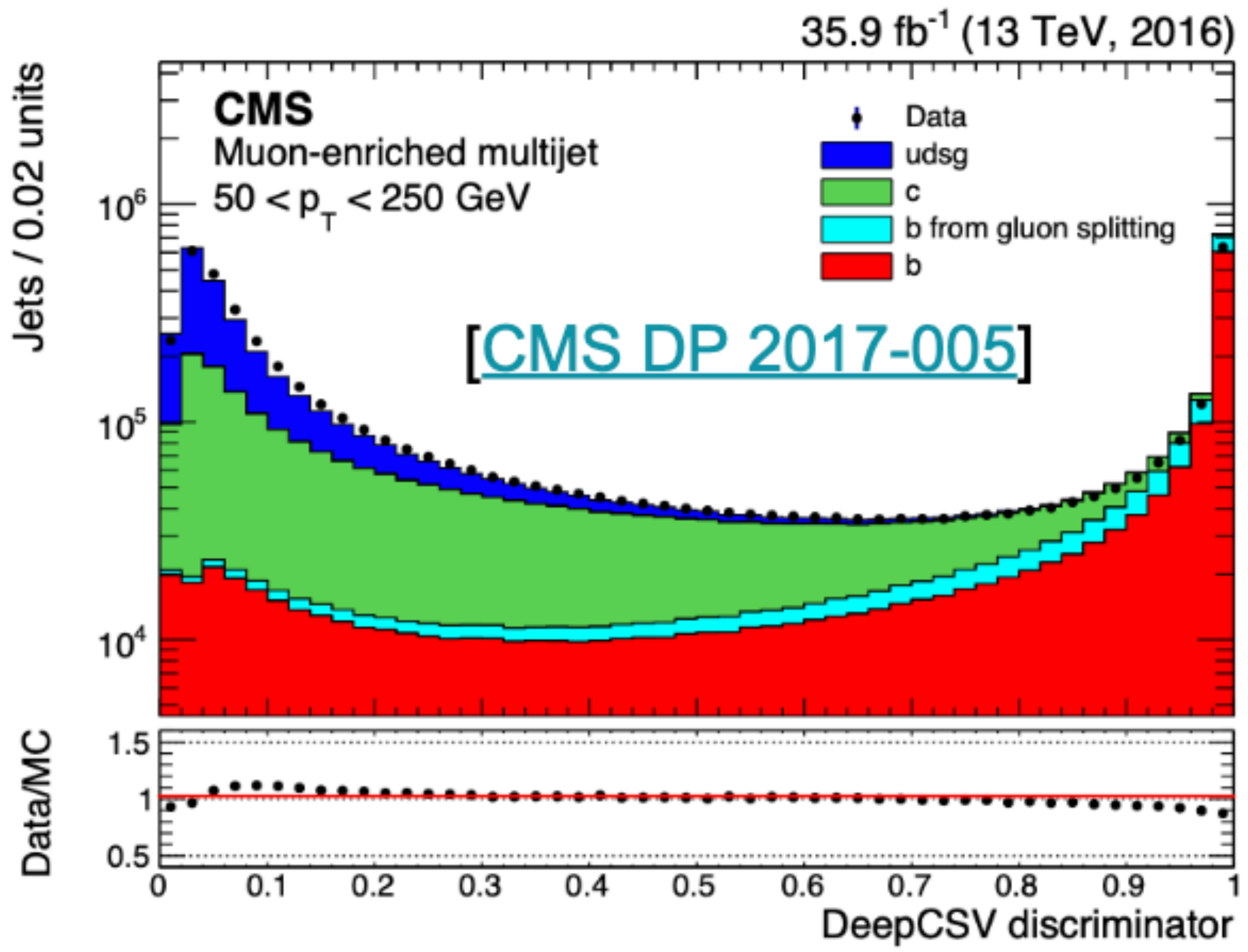
Necessity of SSL in LHC Physics

- Simulations don’t model the data perfectly: need a way to directly train on data
- It will be even harder and more computationally expensive to produce high-quality simulations for High Luminosity LHC (1803.04165)



Input variable

Mismodelling propagates to discriminant



Output discriminant

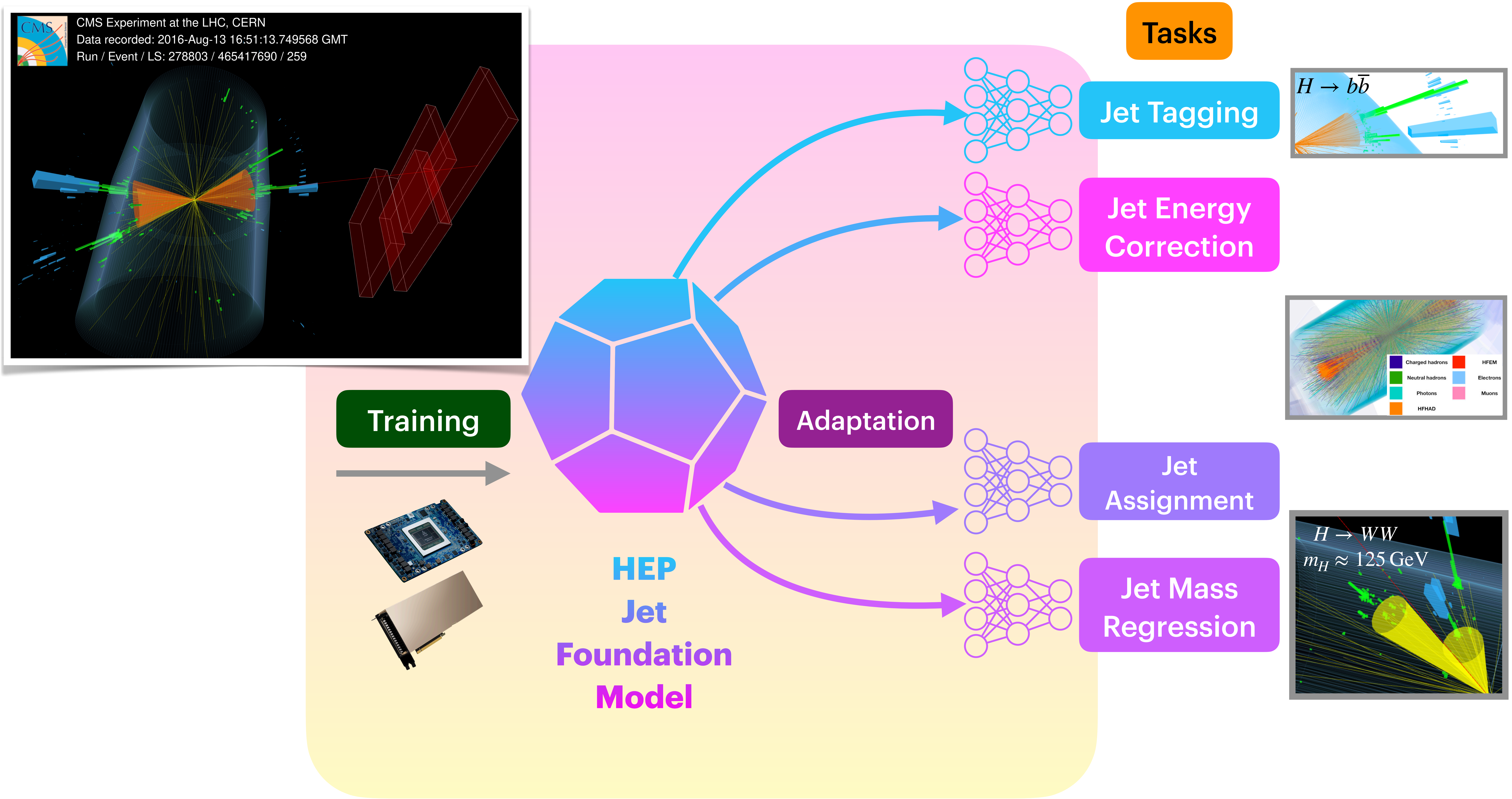
First Goal of the Project

- To show that we can leverage SSL to learn powerful, generic, and transferable features directly from vast unlabeled data.

Current workflow using only Supervised Learning

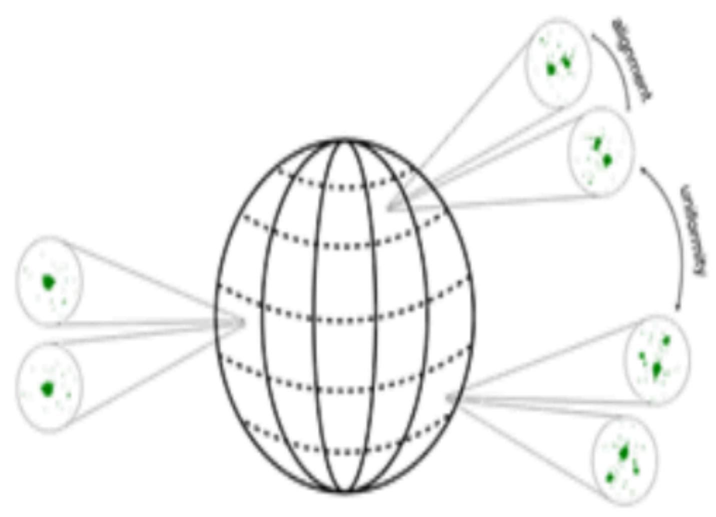
Workflow incorporating SSL

Toward Foundation Model



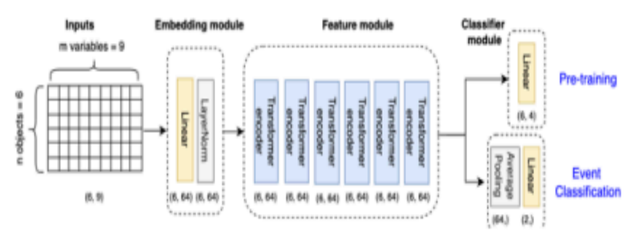
Towards Foundation Model in HEP

Contrastive Learning: Symmetry Augmentation



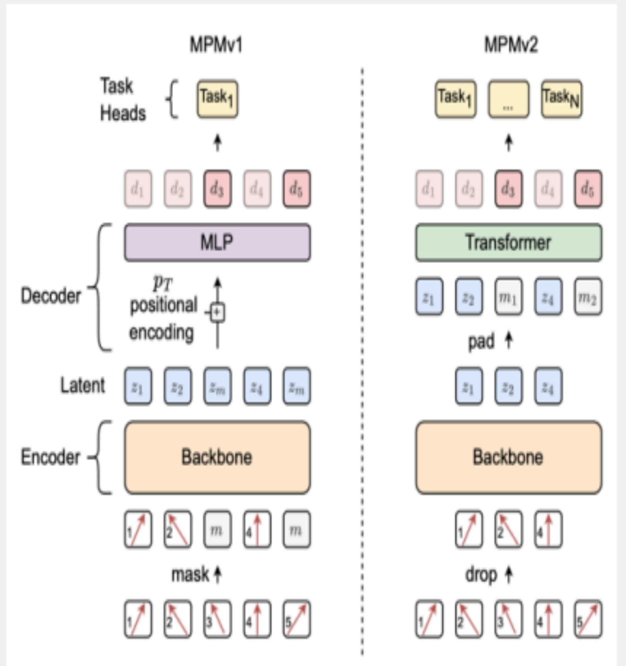
Dillon, Kasieczka, Olischlager Plehn, Sorrenson, Vogel, 2108.04253

Masked Particle Type Prediction



Kishimoto, Morinaga, Saito Tanaka, 2312.06909

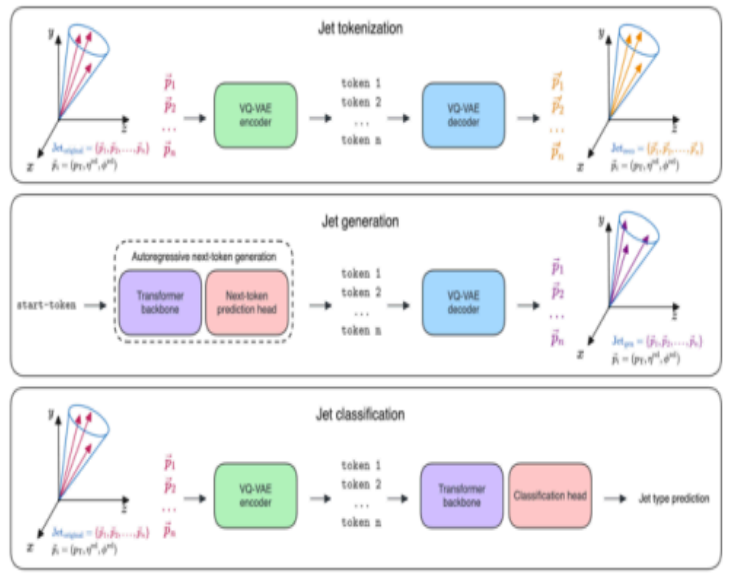
Masked Particle Modeling



Golling, Heinrich, **MK**, Klein, Leigh, Osadchy, Raine, 2401.13537

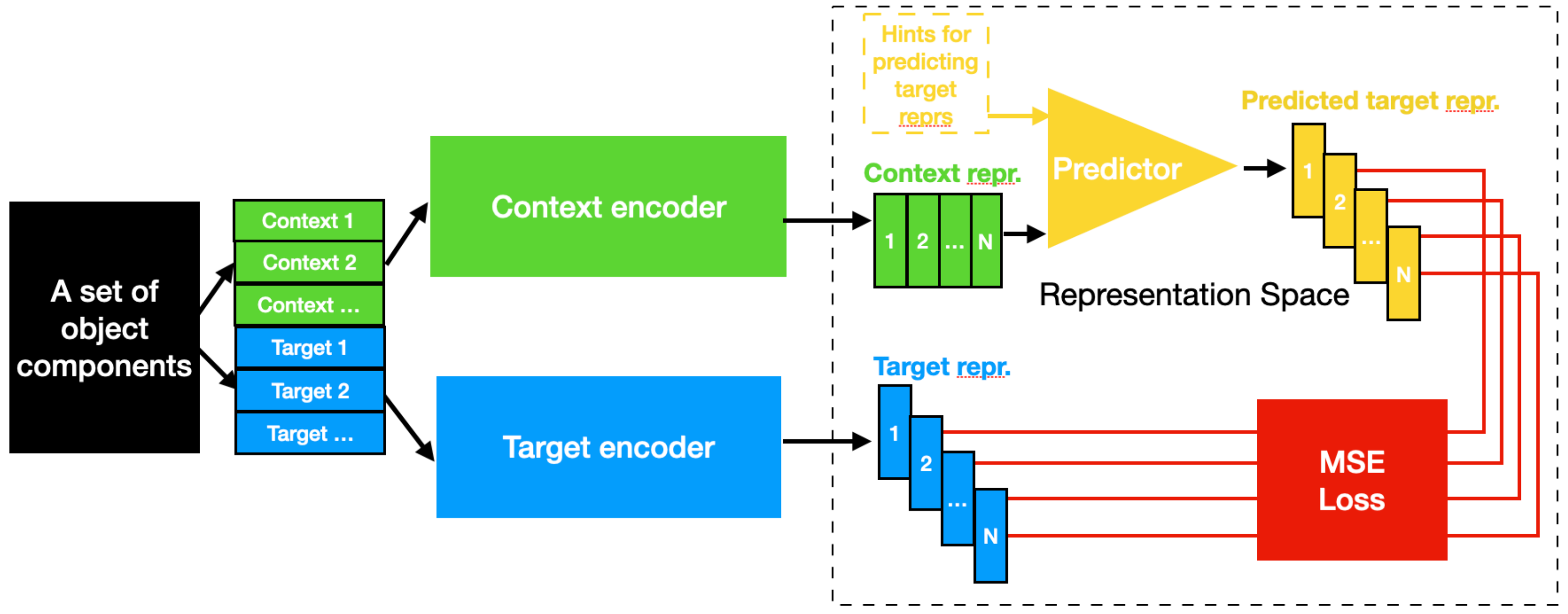
Leigh, Klein, Charton, Golling, Heinrich, **MK**, Ochoa, Osadchy, 2409.12589

Next Token Predictoin



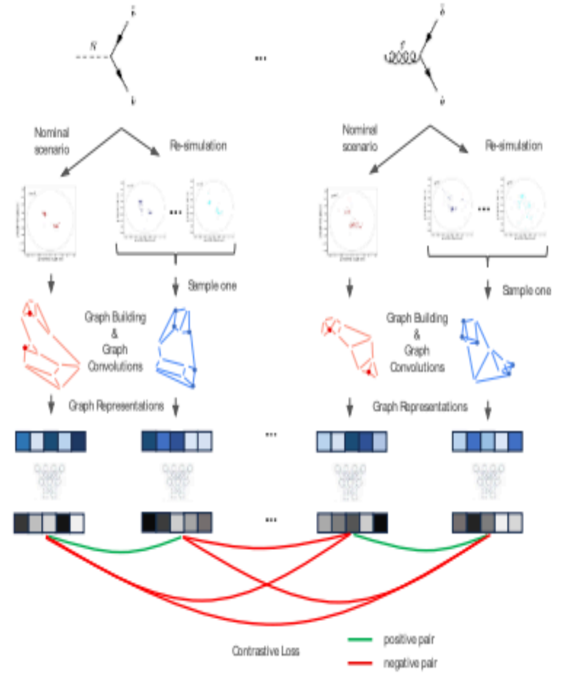
Birk, Hallin, Kasieczka, 2403.05618

J-JEPA



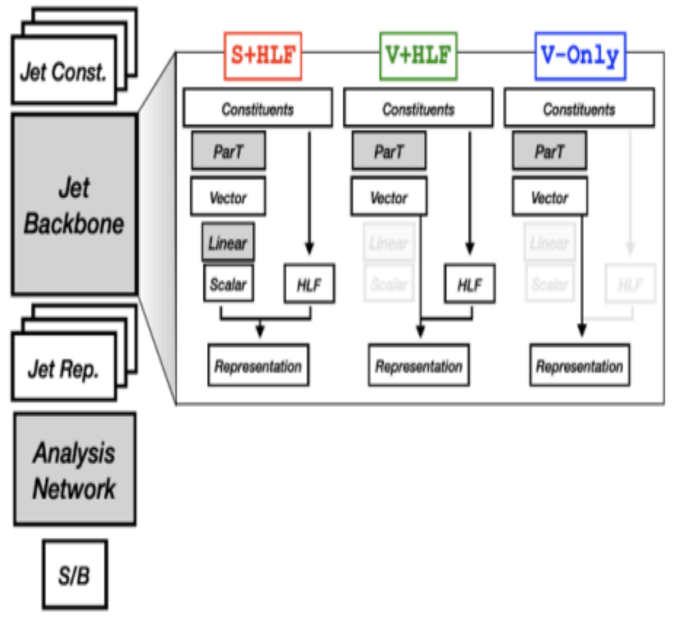
Katel, Li, Zhao, et al.
https://indico.cern.ch/event/1386125/contributions/6083379/

Contrastive Learning: Re-Simulation



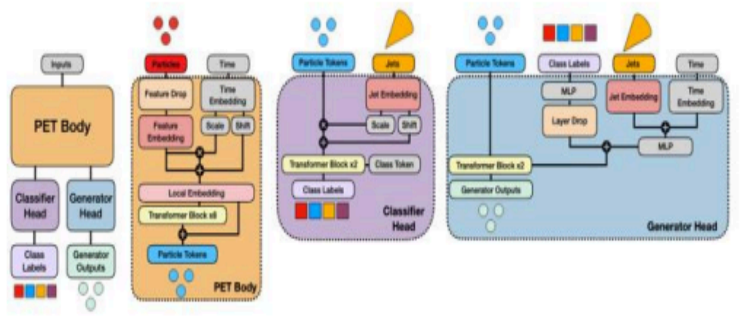
Harris, **MK**, Krupa, Maier, Woodward, 2403.07066

Supervised Pre-training and Joint Optimization



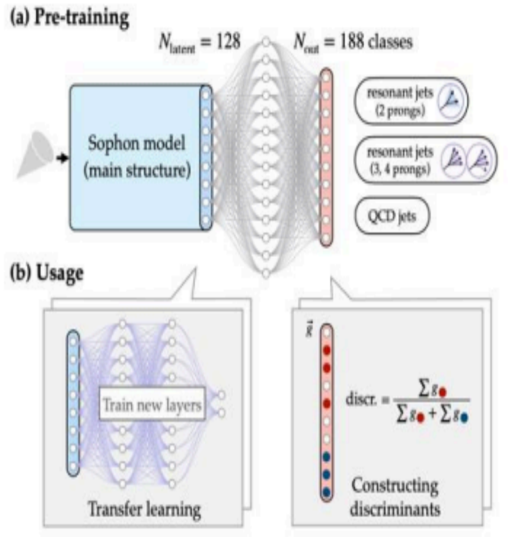
Vigl, Hartman, Heinrich, 2401.13536

Supervised Classification and Generation



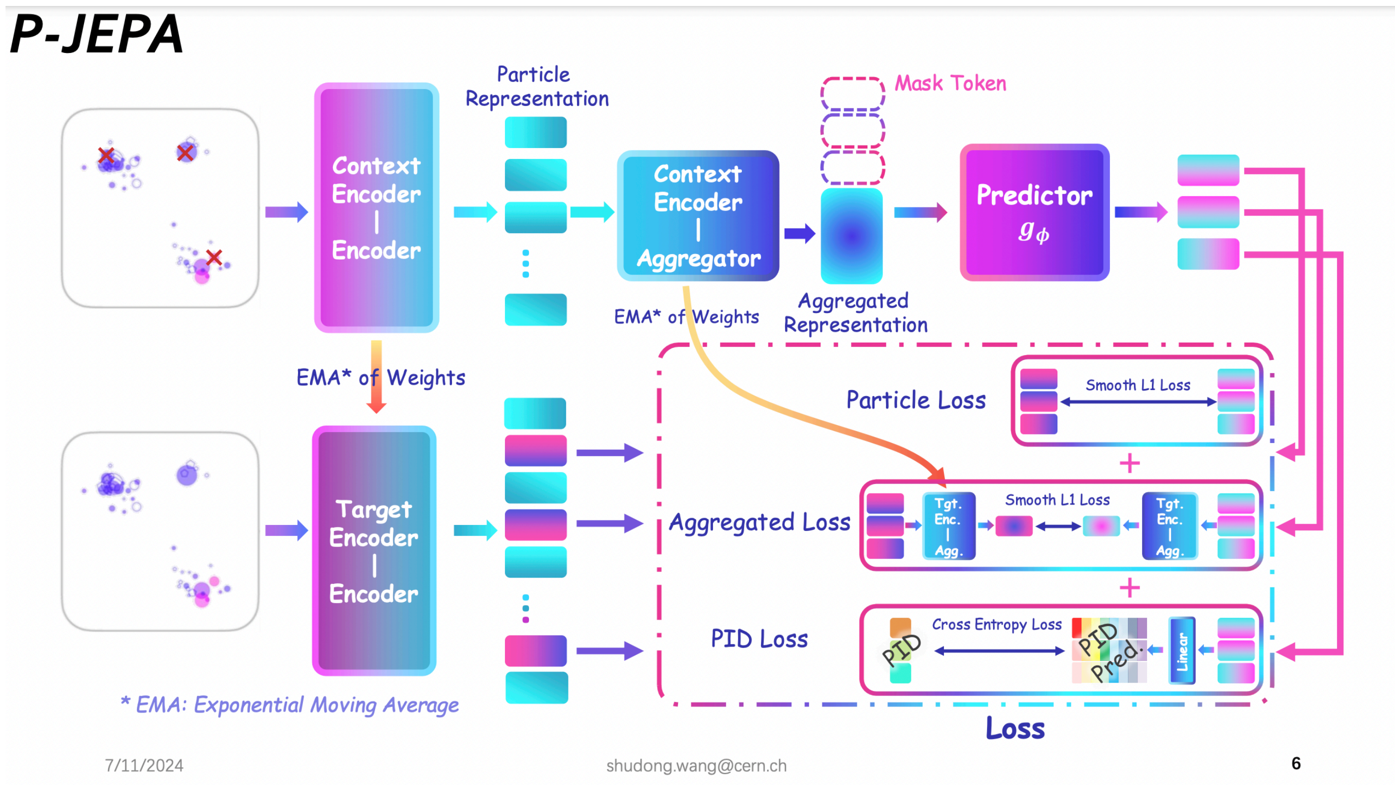
Mikuni, Nachman 2404.16091

Large-Scale Fine-Grained Classification



Li, Li, et al. 2405.12972

...

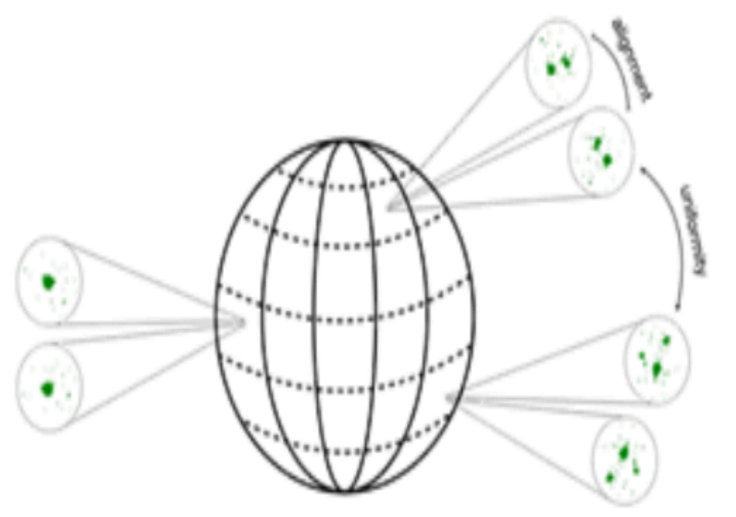


https://indico.cern.ch/event/1386125/contributions/6139666/

Credit: This slide is copied from Michael Kagan's talk in the FM Mini Workshop in October 2024

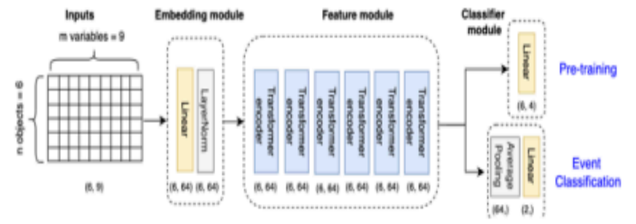
Towards Foundation Model in HEP

Contrastive Learning: Symmetry Augmentation



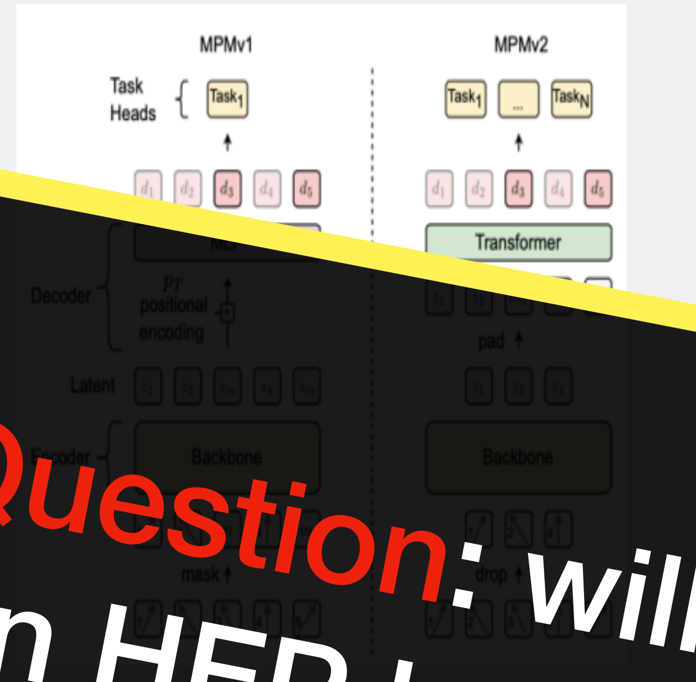
Dillon, Kasieczka, Olischlager
Plehn, Sorrenson, Vogel, 2108.04253

Masked Particle Type Prediction



Kishimoto, Morinaga, Saito
Tanaka, 2312.06909

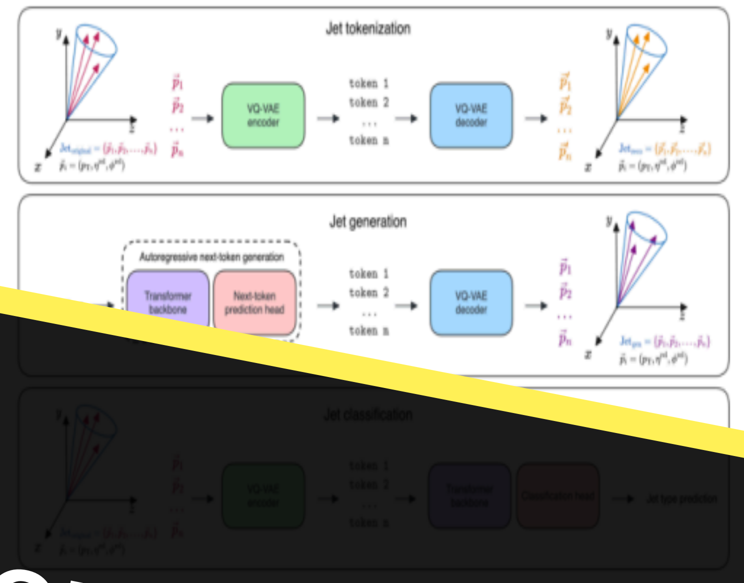
Masked Particle Modeling



Golling, Heinrich, ...
Leigh, Osadchy, Raine,
2401.13537

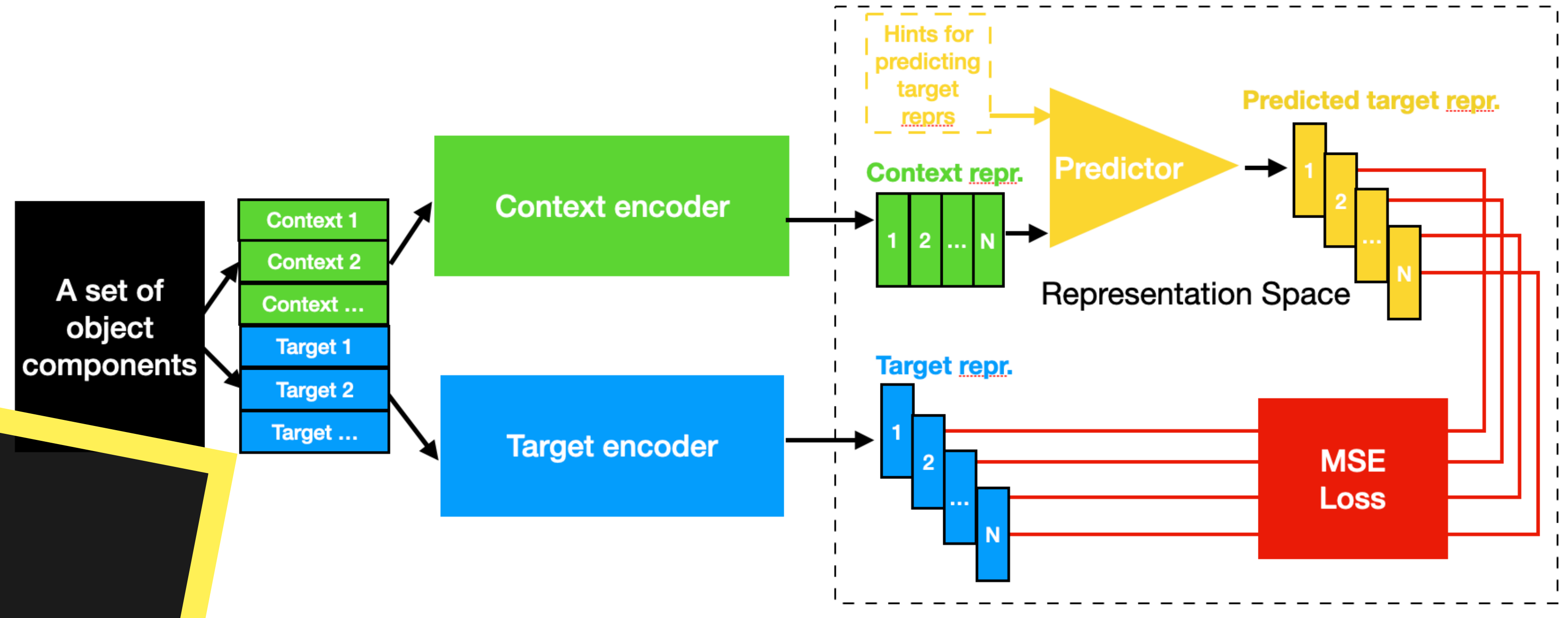
2409.12589

Next Token Predictoin



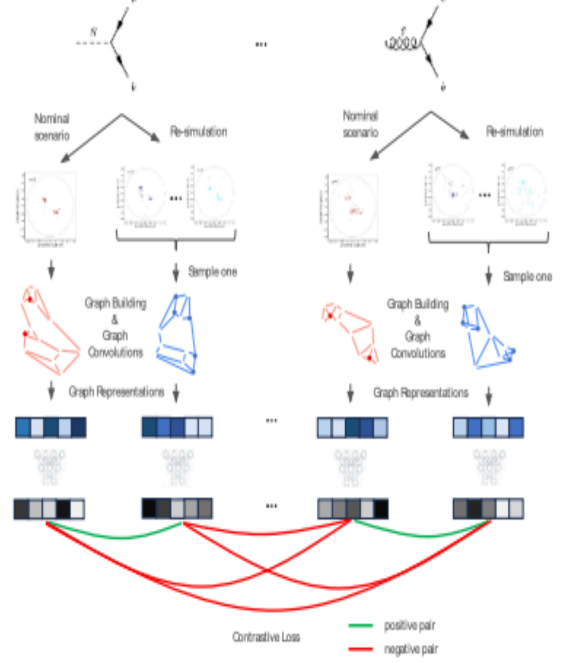
Question: will foundation models in HEP benefit from large scale pretraining?

J-JEPA



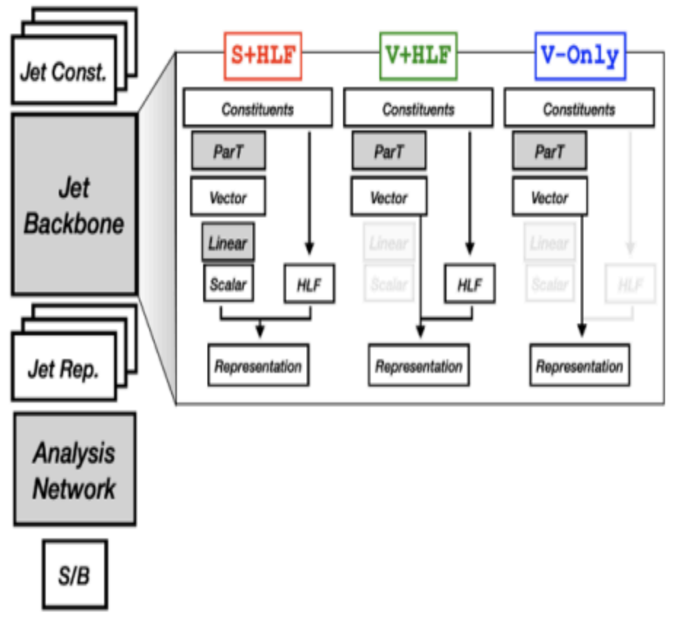
Katel, Li, Zhao, et al.
https://indico.cern.ch/event/1386125/contributions/6083379/

Contrastive Learning: Re-Simulation



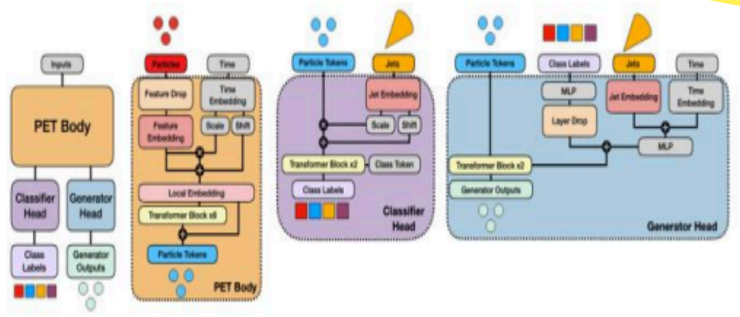
Harris, MK, Krupa, Maier, Woodward, 2403.07066

Supervised Pre-training and Joint Optimization



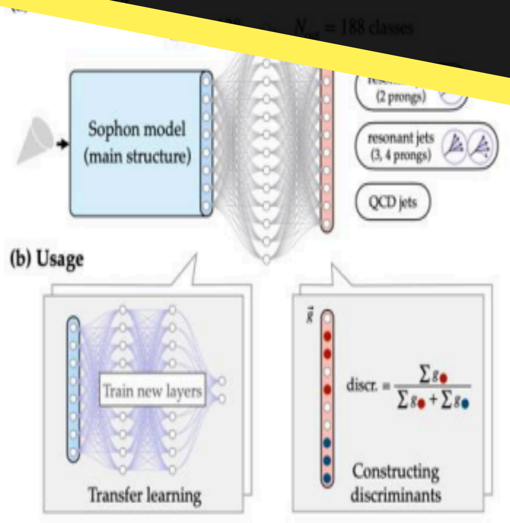
Vigl, Hartman, Heinrich, 2401.13536

Supervised Classification and Generation



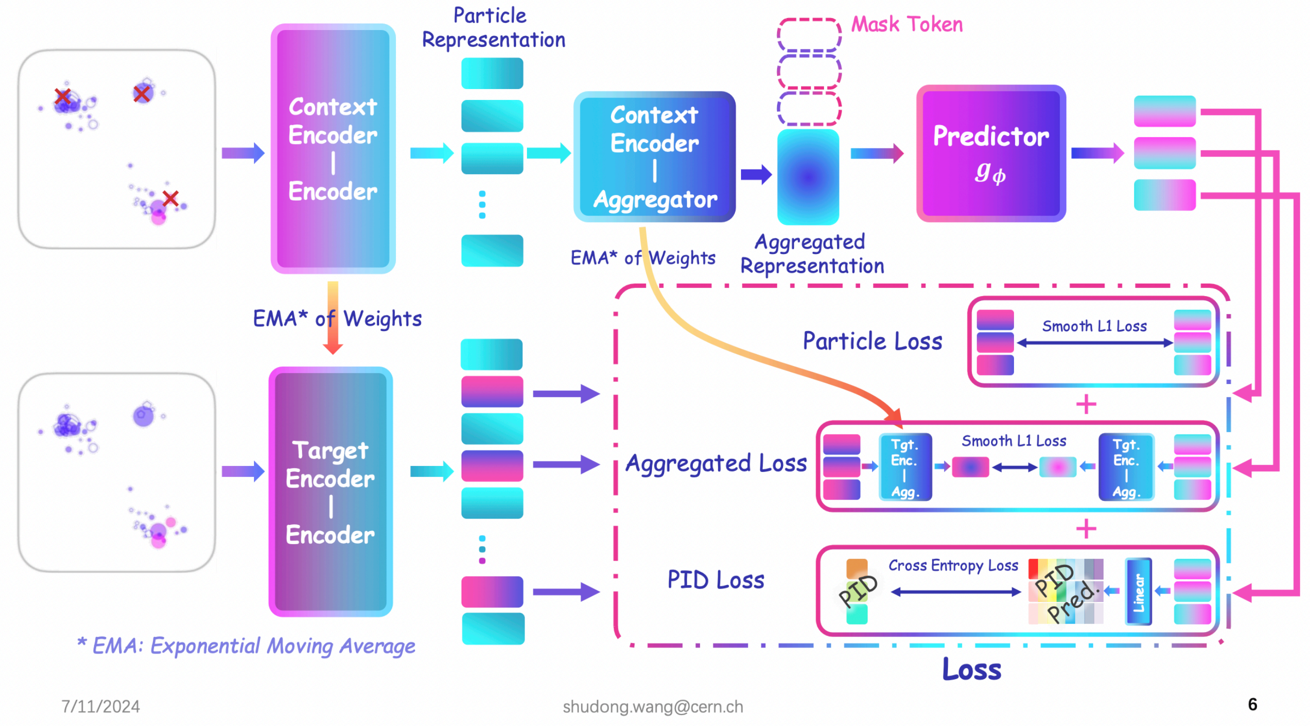
Mikuni, Nachman 2404.16091

Large Scale Fine-Grained Classification



Li, Li, et al. 2405.12972

P-JEPA

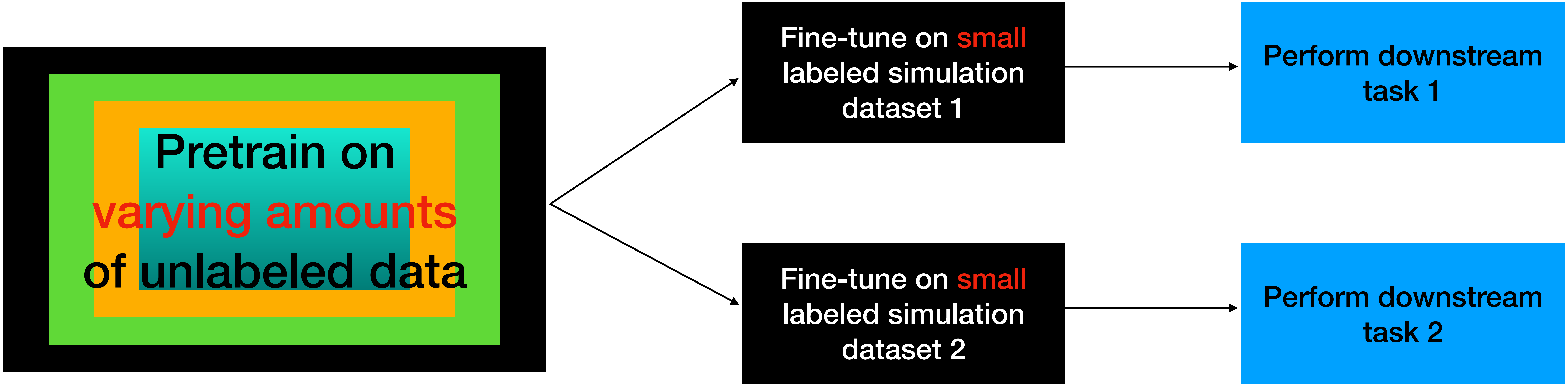


https://indico.cern.ch/event/1386125/contributions/6139666/

Credit: This slide is copied from Michael Kagan's talk in the FM Mini Workshop in October 2024

Primary Goal of the Project

- Focus on studying the effect of **scaling up** the sizes of pretraining datasets on the performance of foundation model.

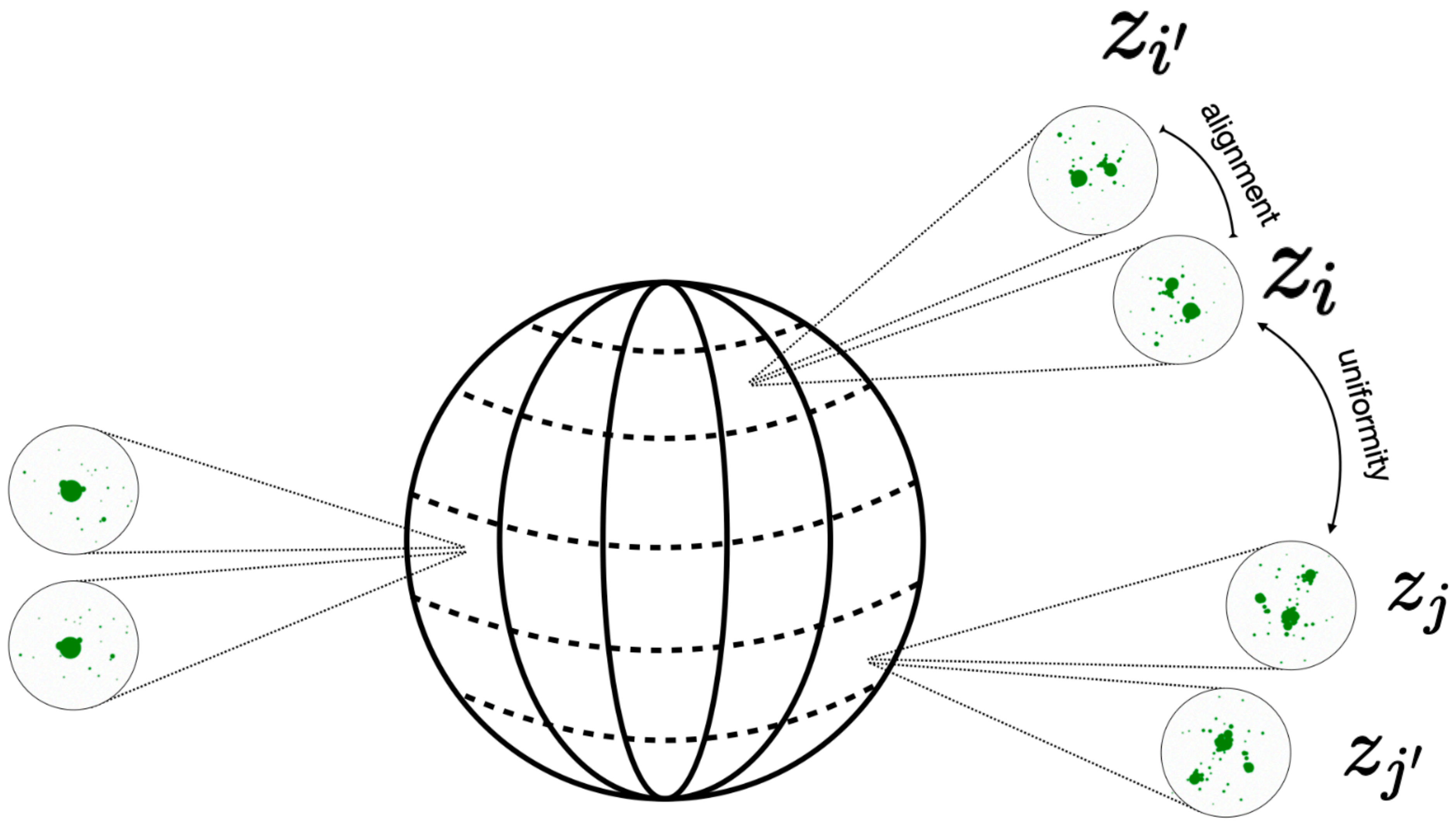


Outline

- Toward Foundation Model in HEP
- Goals of the Project
- Intro to JetCLR
- Transfer Learning: from JetClass to Top Tagging
- Scaling up pretraining dataset size
- Some technical details
  - Classification head for finetuning: MLP vs Linear Projection
  - Techniques to speed up training
- Ongoing and Future work

Intro to JetCLR

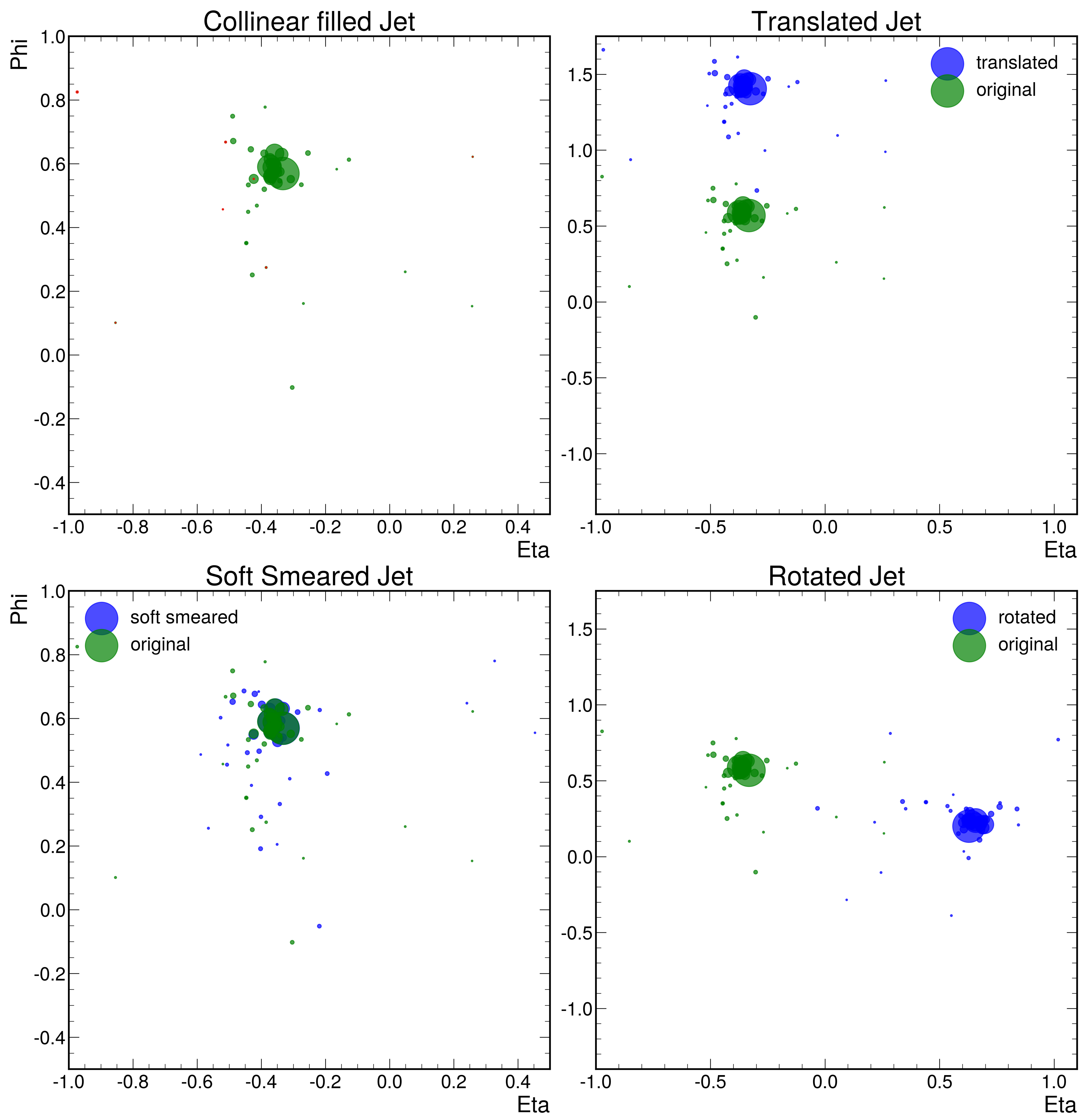
SimCLR loss



$$\mathcal{L}_i = -\log \frac{e^{s(z_i, z_i')/\tau}}{\sum_{j\neq i \in \text{batch}} \left[e^{s(z_i,z_j)/\tau} + e^{s(z_i,z_j')/\tau} \right]}$$

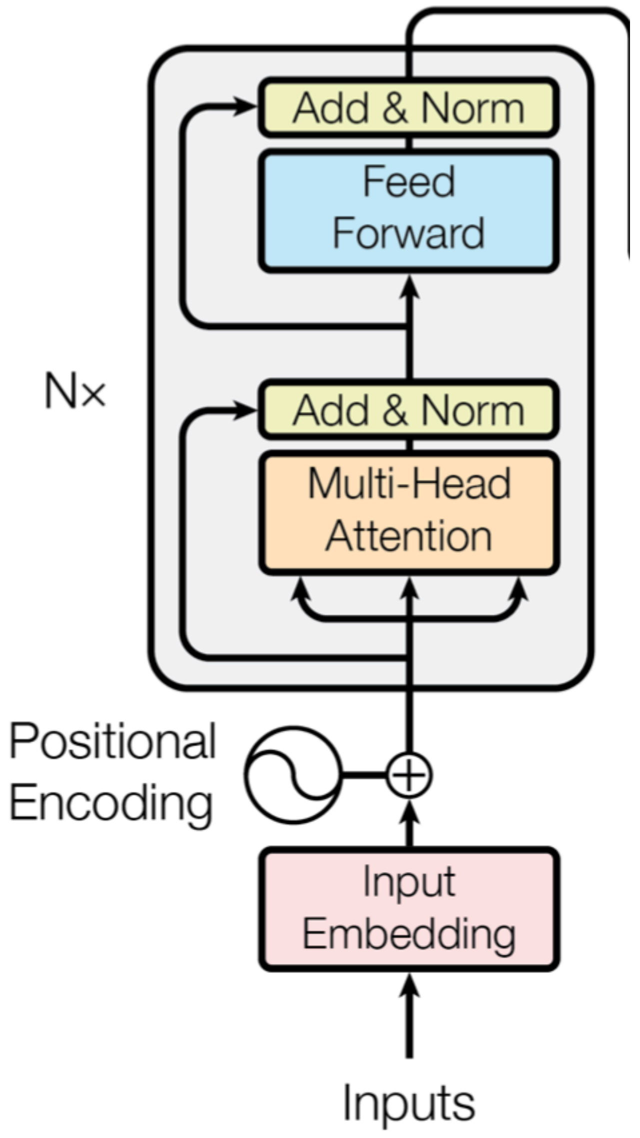
2108.04253

Augmentations



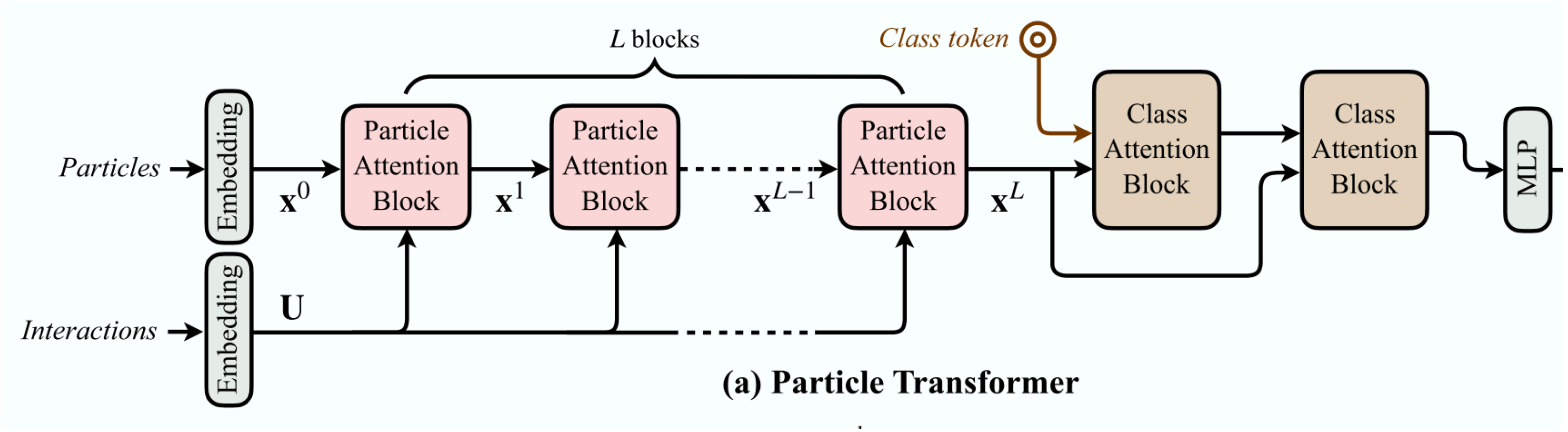
Model Architecture for encoder

- Started with a simple Transformer encoder
- Working on switching to more advanced architectures such as Particle Transformer



Transformer Encoder

1706.03762



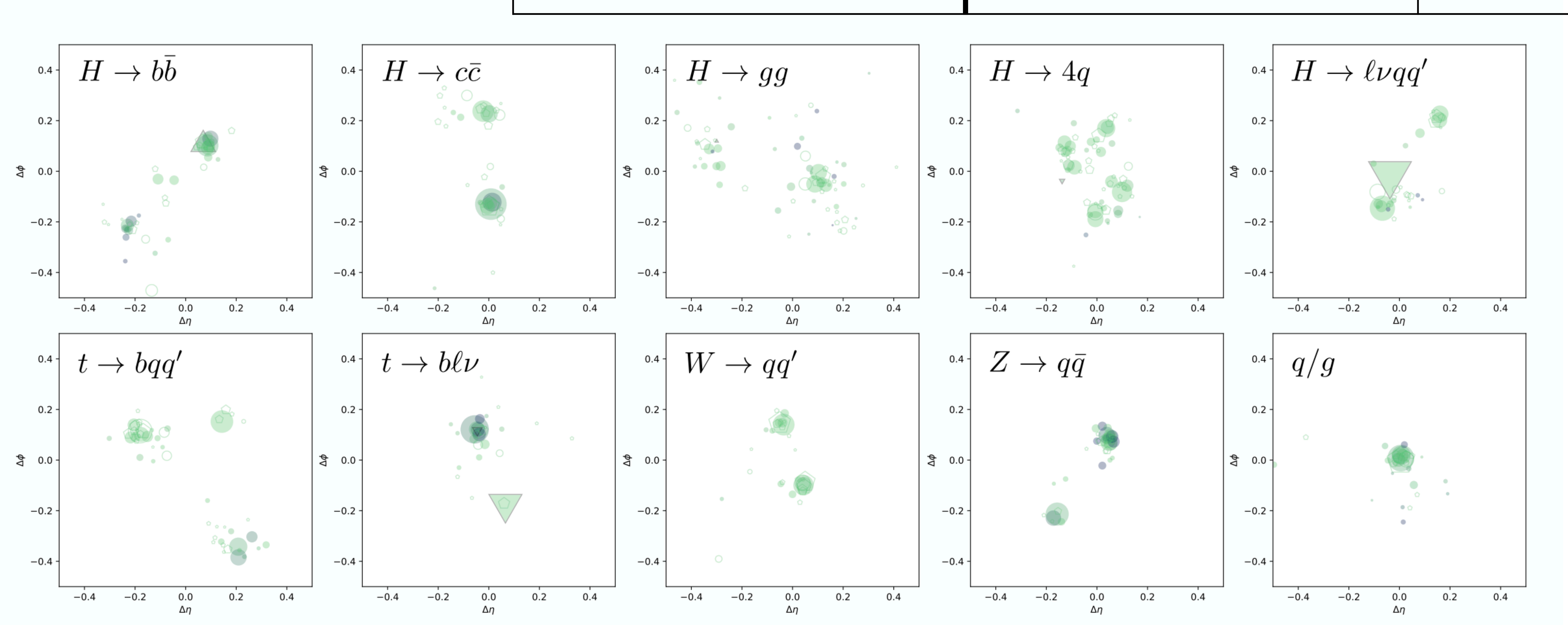
Particle Transformer

2202.03772

Datasets

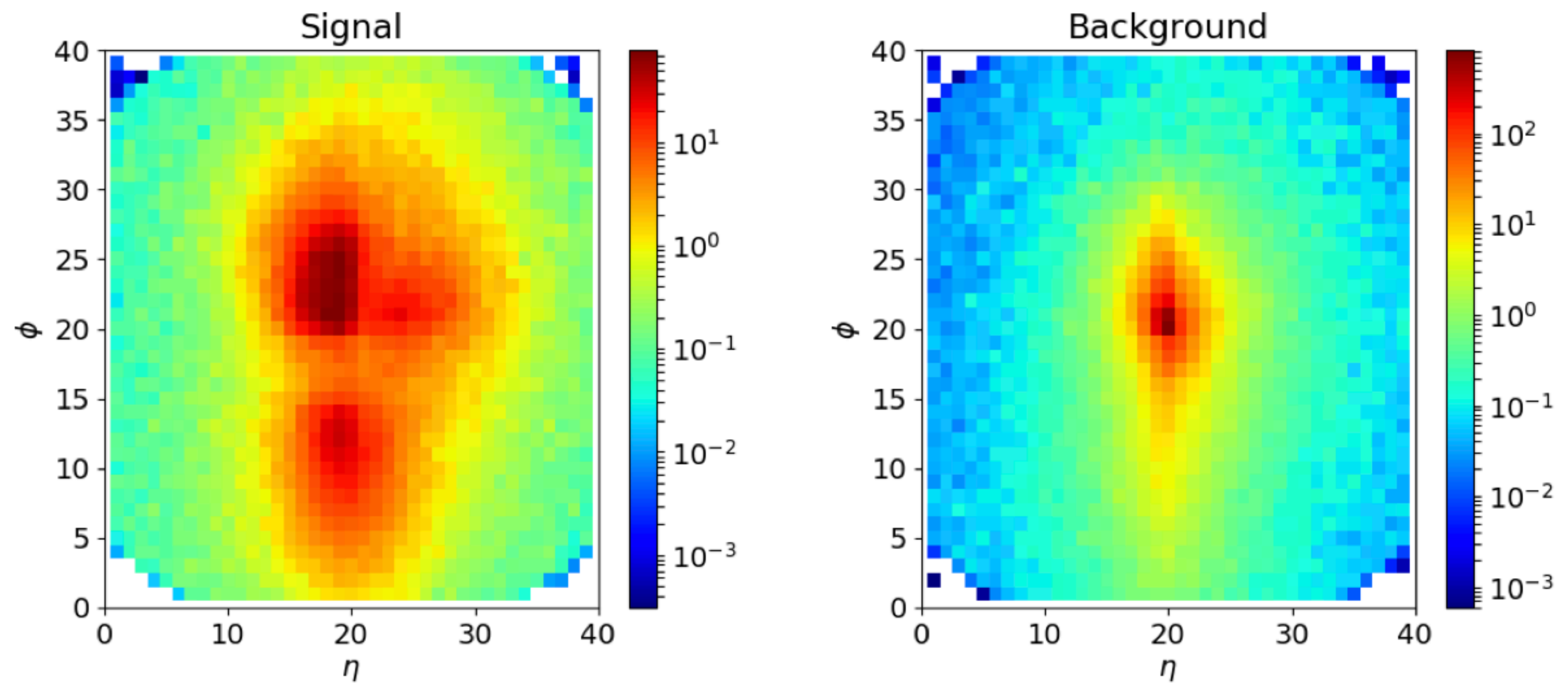
JetClass for unlabeled pretraining, Top Tagging for labeled finetuning

Dataset name	Size	Description	Role in transfer learning
JetClass Dataset	100 Million Jets	Contains 10 classes of jets	Stand in for unlabeled “data”, use for pretraining
Top Tagging Dataset	1.2 Million Jets	Only Top and QCD jets	Stand in for labeled “simulation”, use for fine-tuning



JetClass Dataset

2202.03772



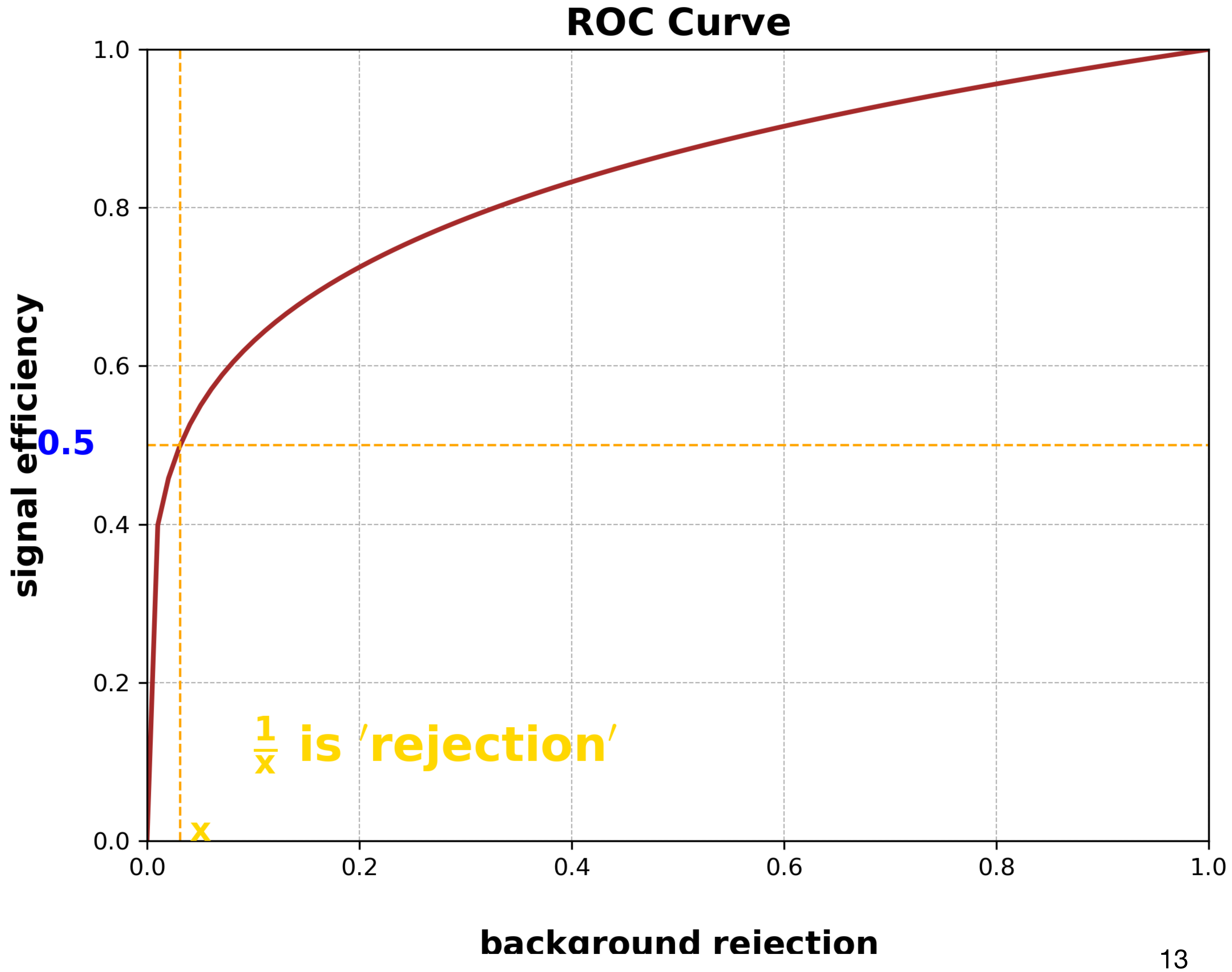
Top Tagging Dataset

1902.09914

Metrics

Accuracy: correctly predicted / total number of samples

Rejection: inverse of background rejection (FPR) at 50% signal efficiency (TPR)



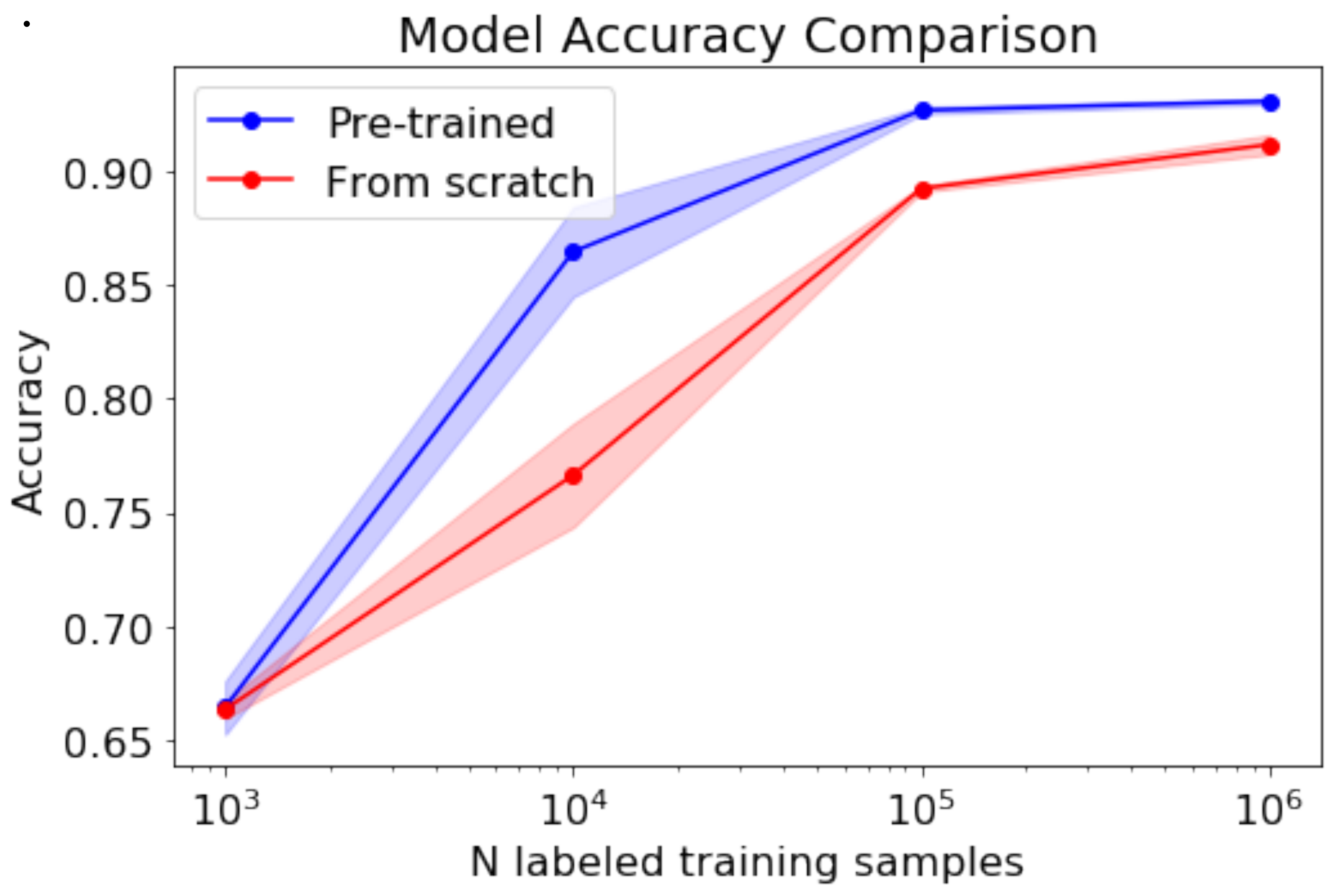
Significance: In a background dominant dataset, how much background you can reject while letting in a certain fraction of signal samples (the more the better)

Pretraining on JetClass and fine-tuning on Top Tagging

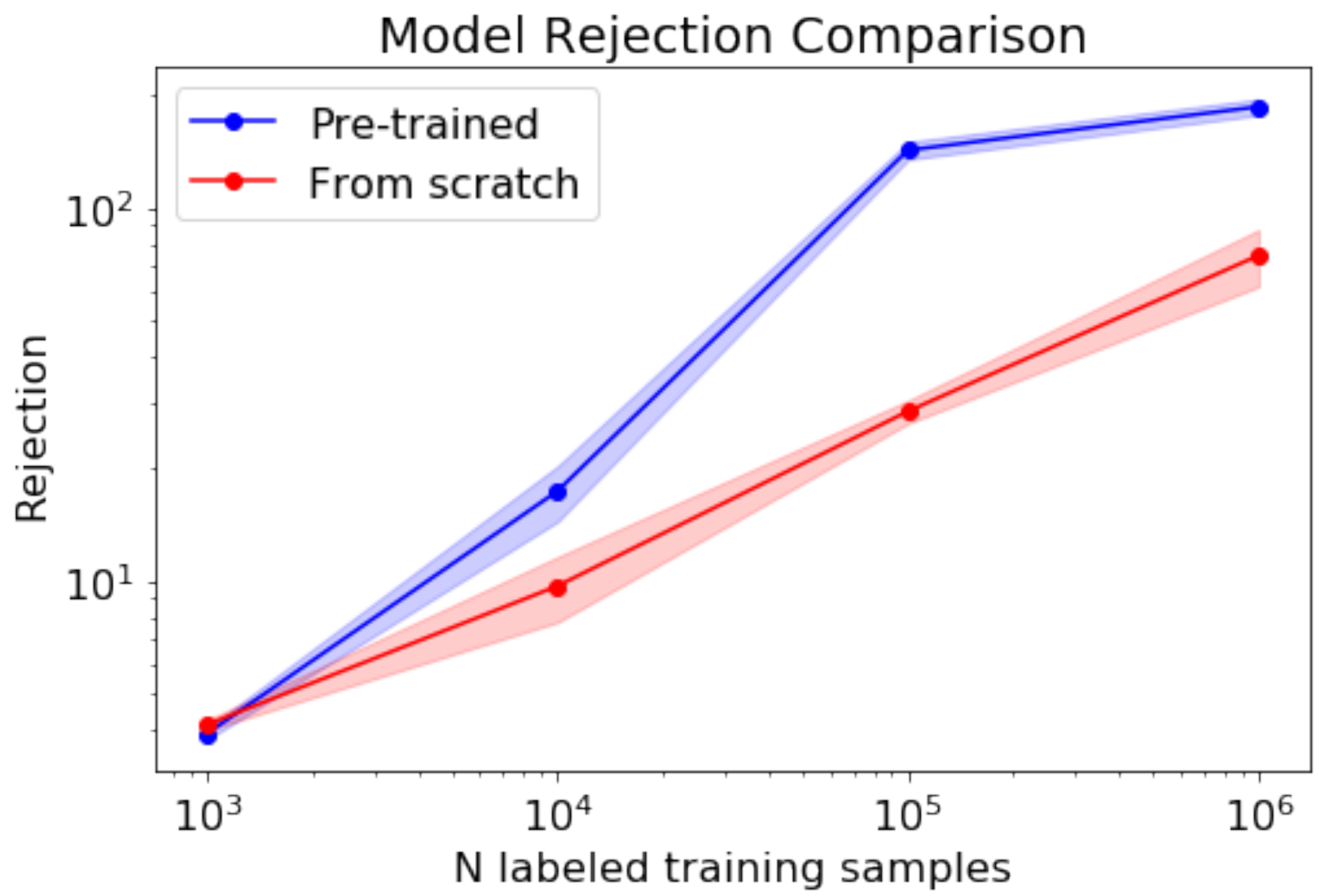
The pre-trained model requires significantly fewer samples to achieve high accuracy and rejection rate: higher data efficiency

- The averages and standard deviations over 5 trainings are shown in solid lines and uncertainty bands, respectively

.



Pretrained: pretrained with 1M jets

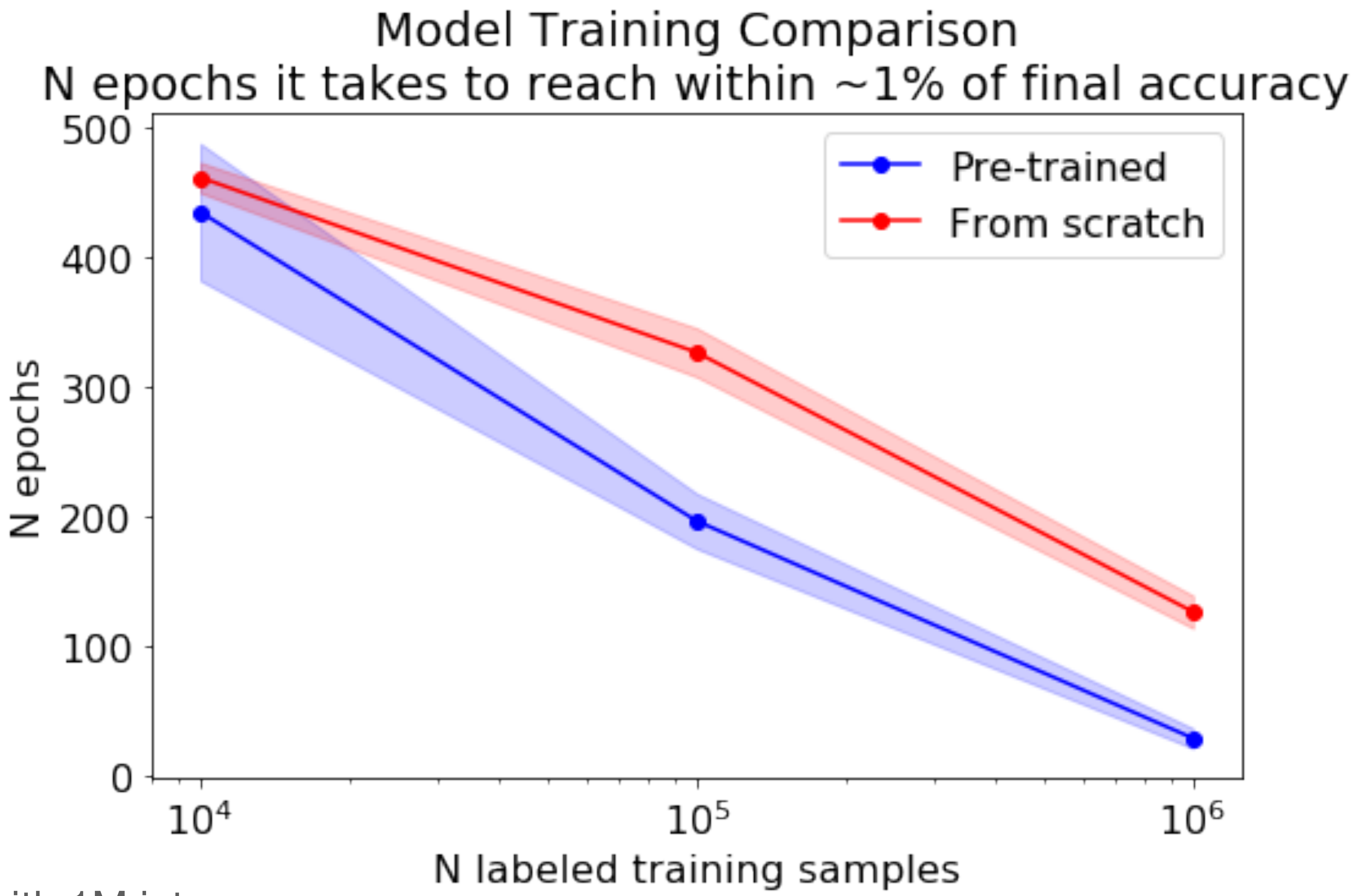


Rejection: inverse of background rejection at 50% signal efficiency

Pretraining on JetClass and fine-tuning on Top Tagging

The pre-trained model converges much faster: higher computational efficiency

- The averages and standard deviations over 5 trainings are shown in solid lines and uncertainty bands, respectively

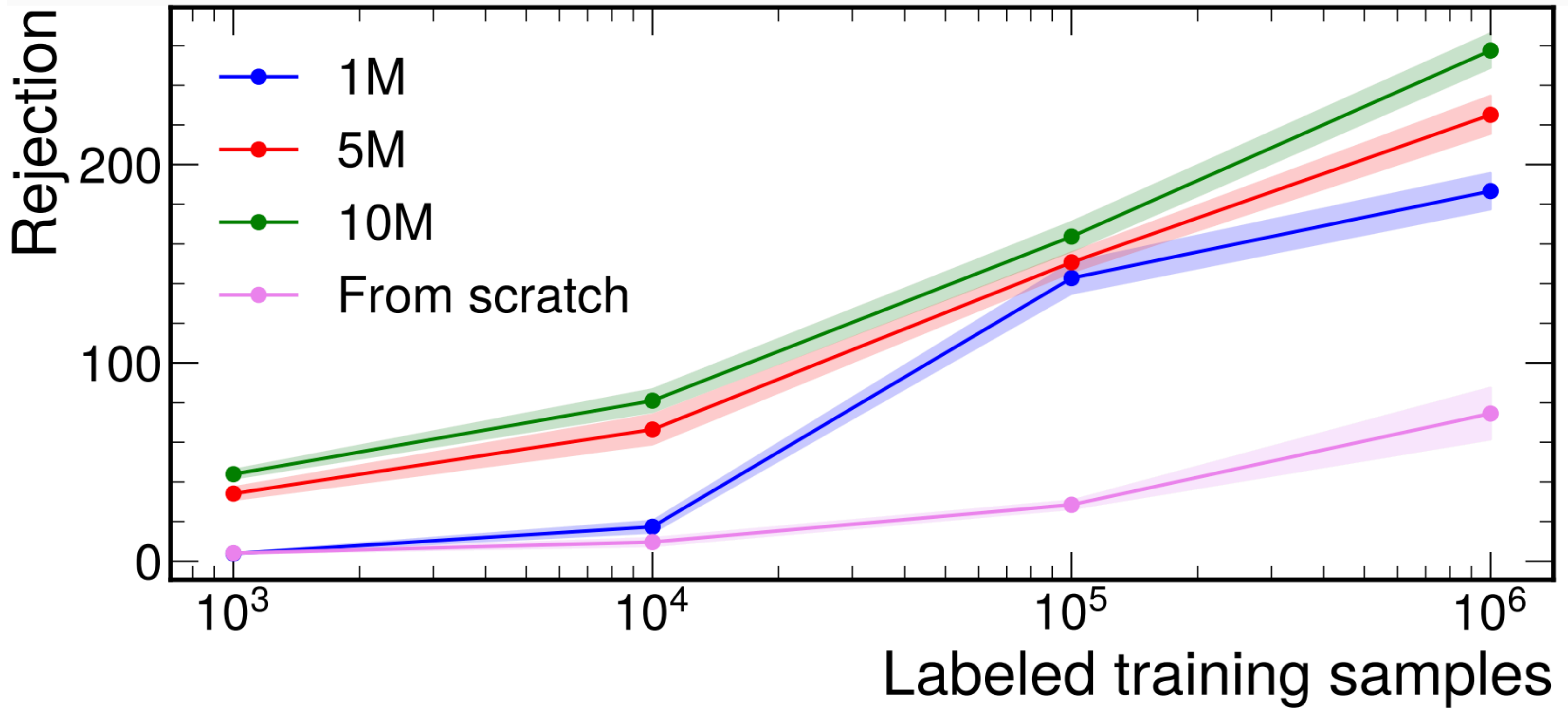


Pretrained: pretrained with 1M jets

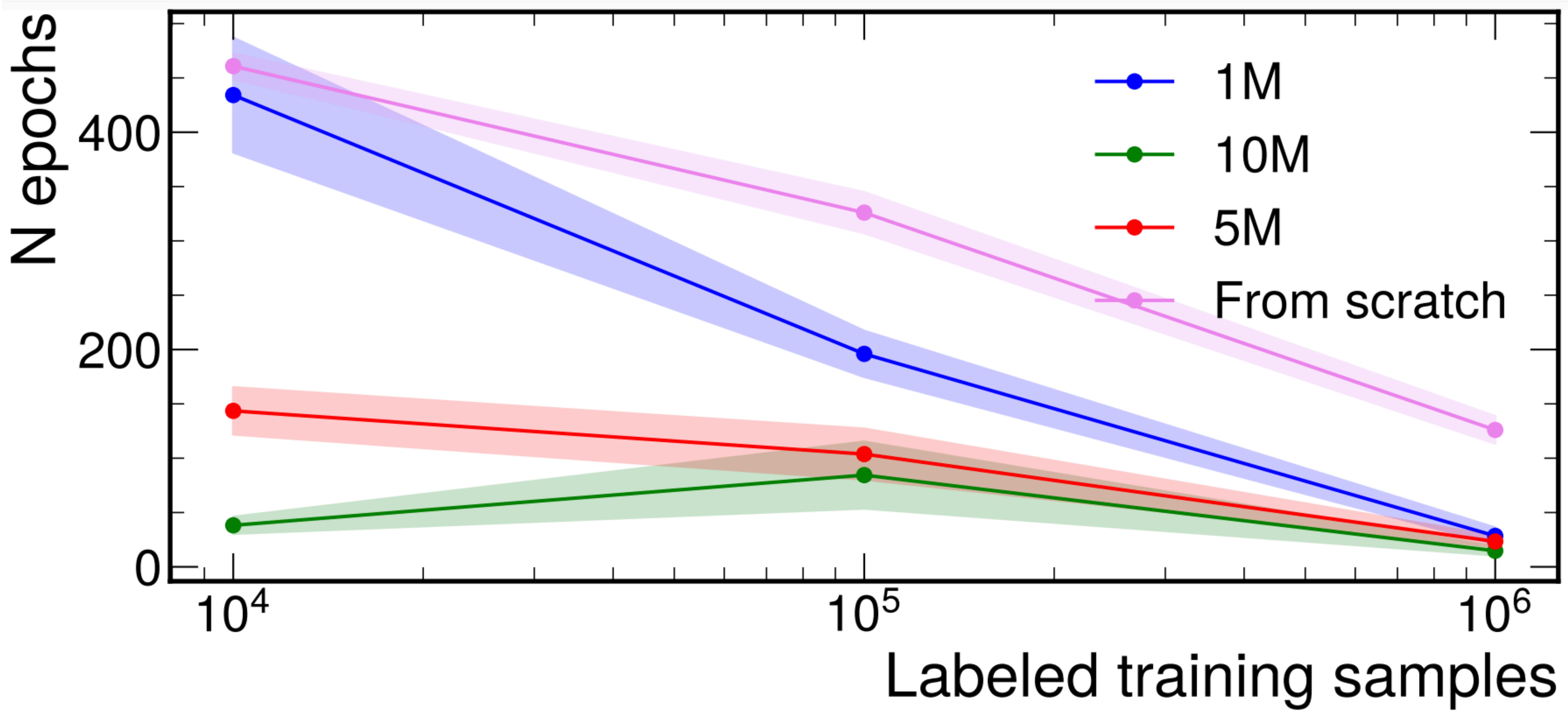
Scaling up pretraining dataset size

By scaling up the pretraining dataset, the model demonstrated enhanced performance and faster convergence: both data and computational efficiency improve as we use larger datasets for pretraining

Background rejection at 50% signal efficiency



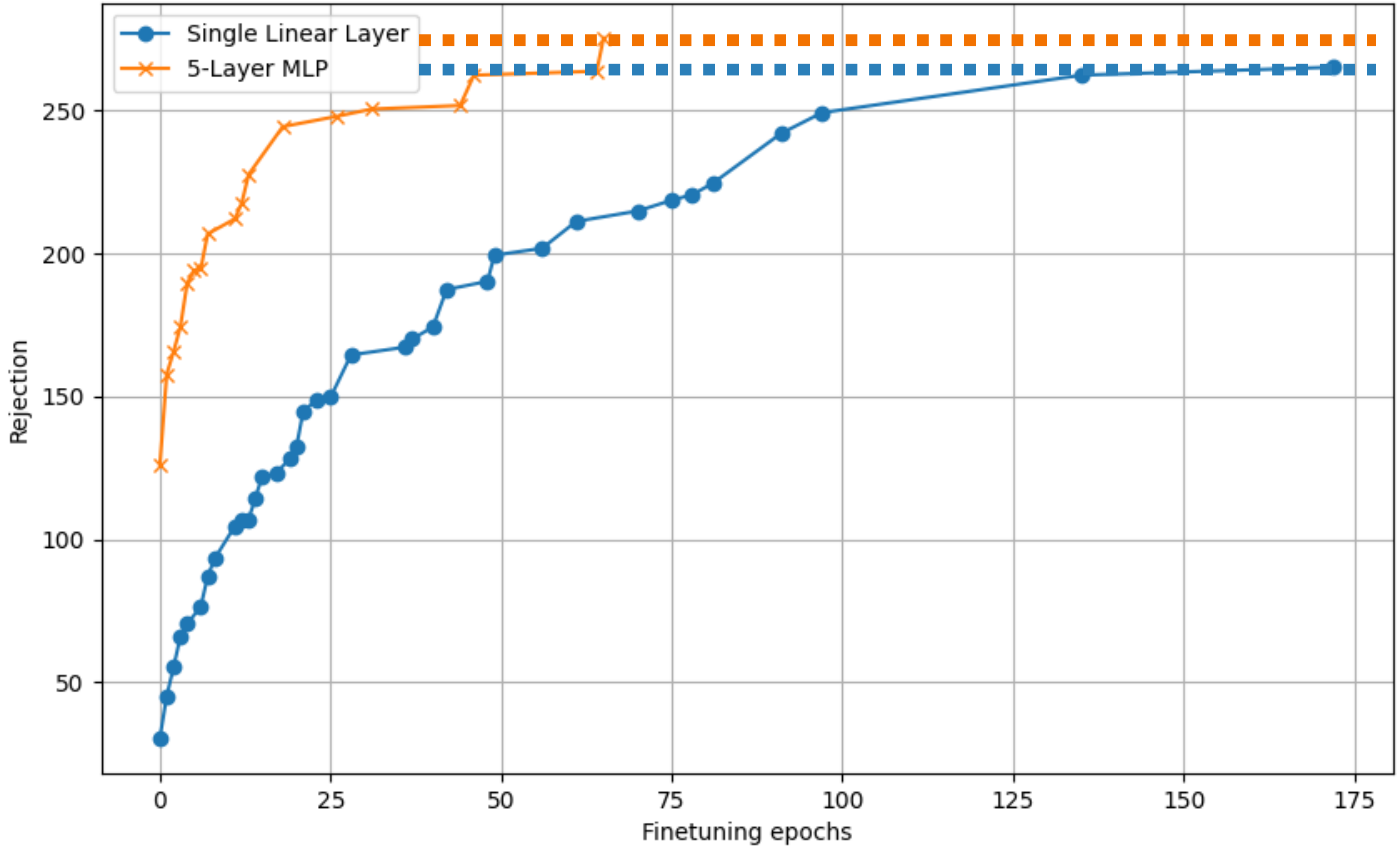
Number of epochs required to reach within 1% of the final accuracy



Rejection: inverse of background rejection at 50% signal efficiency

Classification head for finetuning: MLP vs Linear Projection

Using an MLP with activation helps converge faster, but no significant improvement in performance



Techniques to speed up training

Steps we took to ensure the model finished pretraining within a reasonable amount of time

- Removed unnecessary CPU-GPU synchronizations, especially read-out from GPU for recording losses
- Modified the default model dimensions to be multiples of 8 to make use of CUDA matrix multiplication kernels more efficiently
- Fused point-wise operations into a single CUDA kernel when computing the contrastive loss.
- Utilized the Automatic Mixed Precision (AMP) package
    - Measures to mitigate the numerical instability caused by using AMP in backup.

Conclusion

- Through large-scale pretraining followed by finetuning, our SSL approach has demonstrated
    - **Enhanced data efficiency—**requiring fewer labeled training samples to achieve superior performance compared to the fully supervised approach.
    - **Greater computational efficiency—**enabling the model to converge significantly faster than its fully supervised counterpart.
    - **Both efficiencies increase as the pretraining dataset size increases.**
- This paves the way for the use of unlabeled data in HEP and contributes to a better understanding of the potential of SSL for scientific discovery.

Ongoing and Future work

- Ongoing work
  - Study the effectiveness of more advanced architectures like the ParticleTransformer as the backbone encoder
- Pretrain on JetClass v2, an even larger dataset
- Evaluate on different SSL strategies beyond JetCLR
- Explore other physically motivated augmentations
  - Pairing the two jets from dijet events
  - Using two subjets clustered with smaller radii
  - Using tracks and clusters as two views of the same jet
  - …

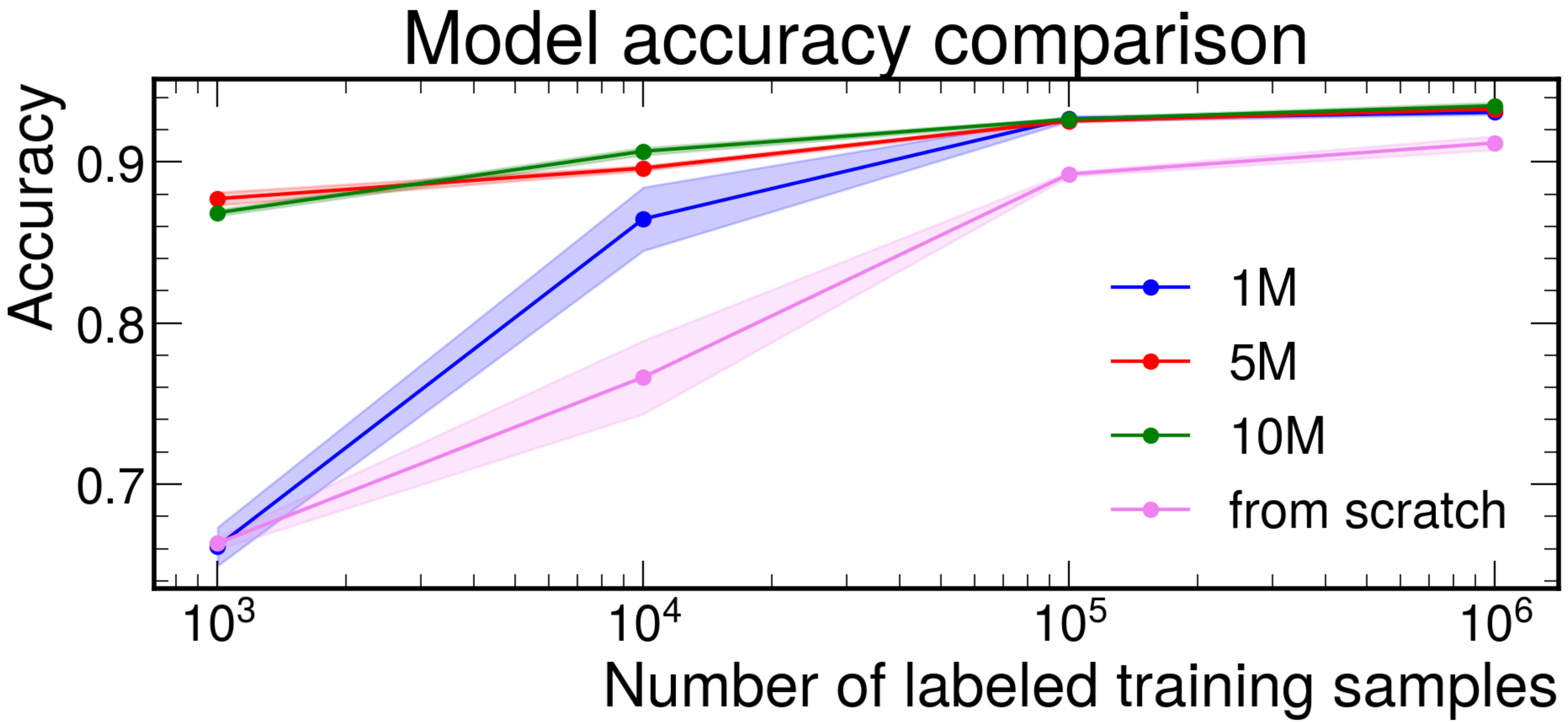
Support

Thank you for listening!

- This work is supported by the National Science Foundation under award number 2117997 (A3D3 Institute), Research Corporation For Science Advancement, the Alfred P. Sloan Foundation, and the U.S. Department of Energy

- This work was performed using the National Research Platform Nautilus HyperCluster supported by NSF

Back Up


Model accuracy comparison
Accuracy
0.9
0.8
0.7
1M
5M
10M
from scratch
10^3
10^4
10^5
10^6
Number of labeled training samples

LHC and Jet Tagging

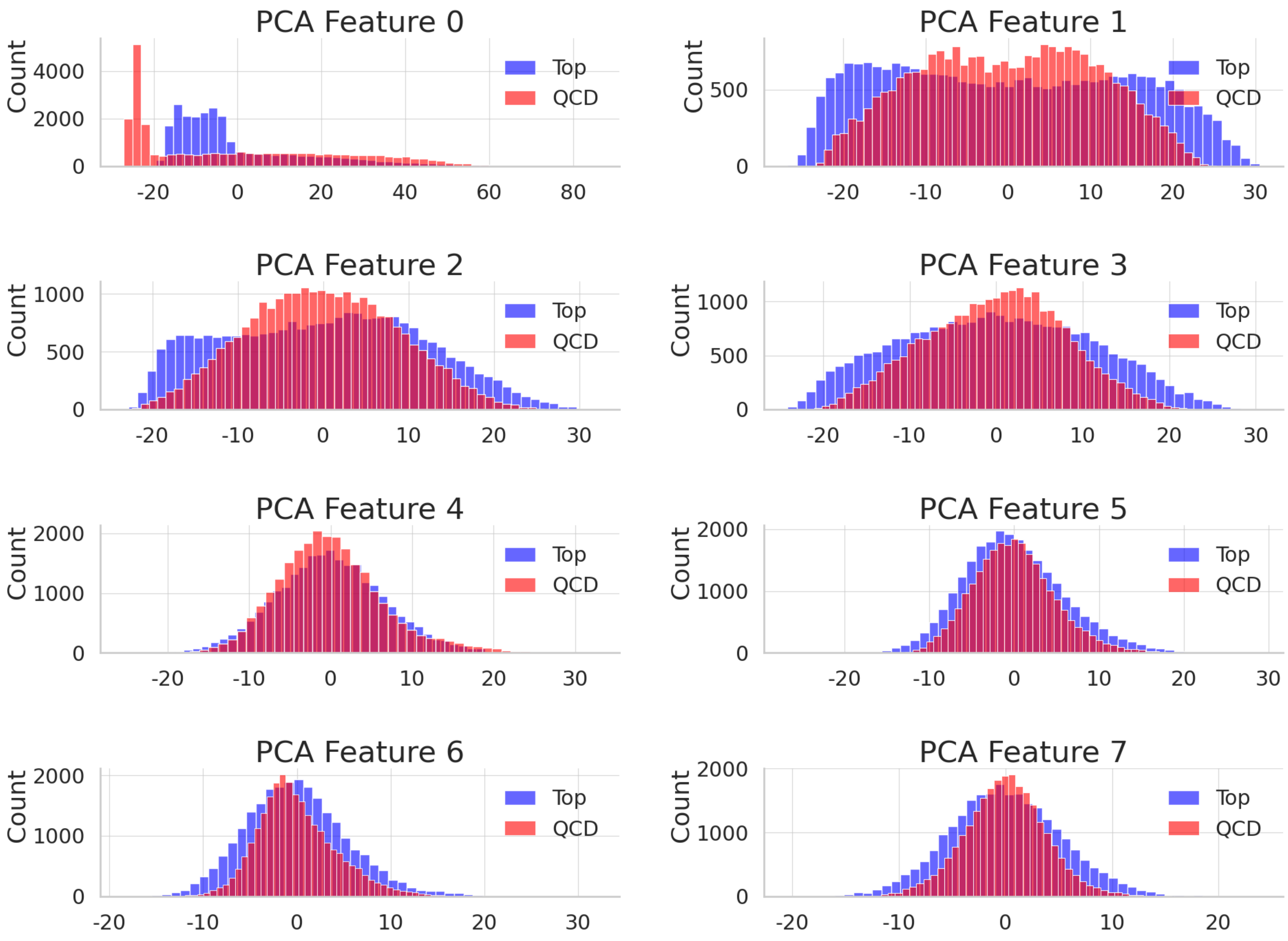
1709.04464

Measures to mitigate the numerical instability caused by using AMP

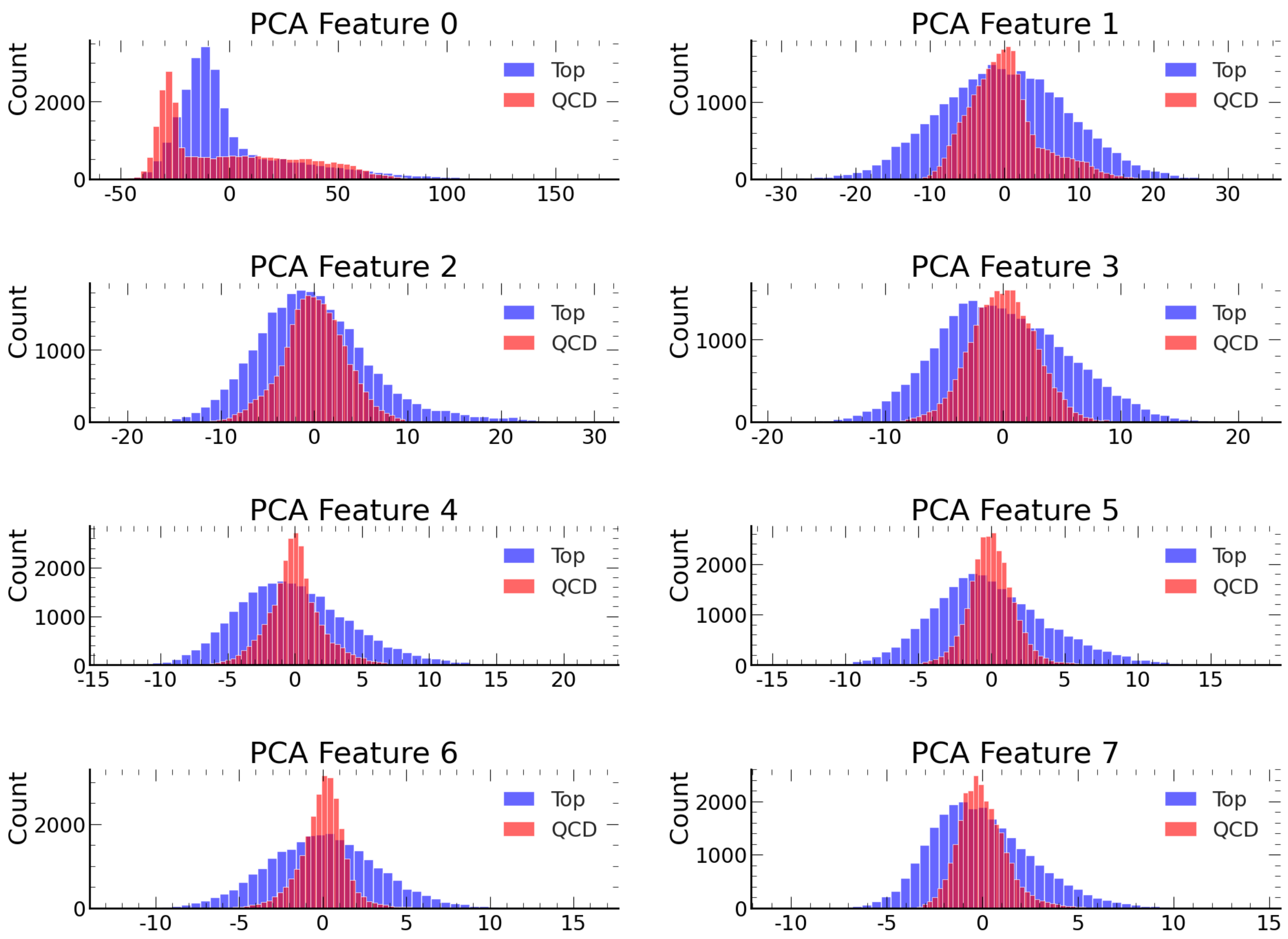
- Monitor loss and gradient values regularly with tensorboard
- Gradient clipping with a maximum norm of 0.1
- Set the ϵ parameter to 10^(−4) in the Adam optimizer.
- Manually run certain parts of the code in full precision

Pretraining on JetClass and fine-tuning on Top Tagging

The pre-trained model shows a much clearer separation between signal and background



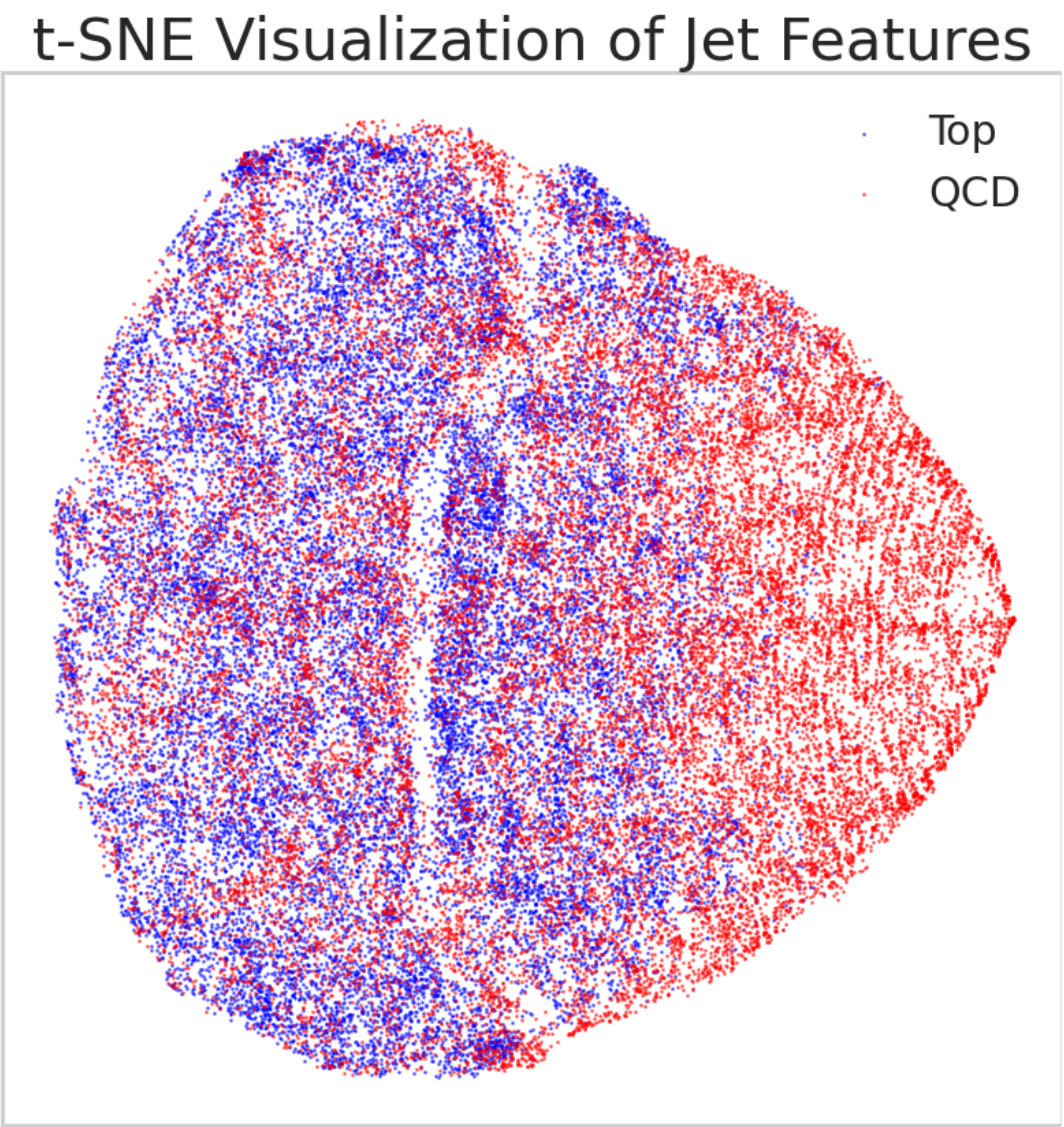
Trained from scratch



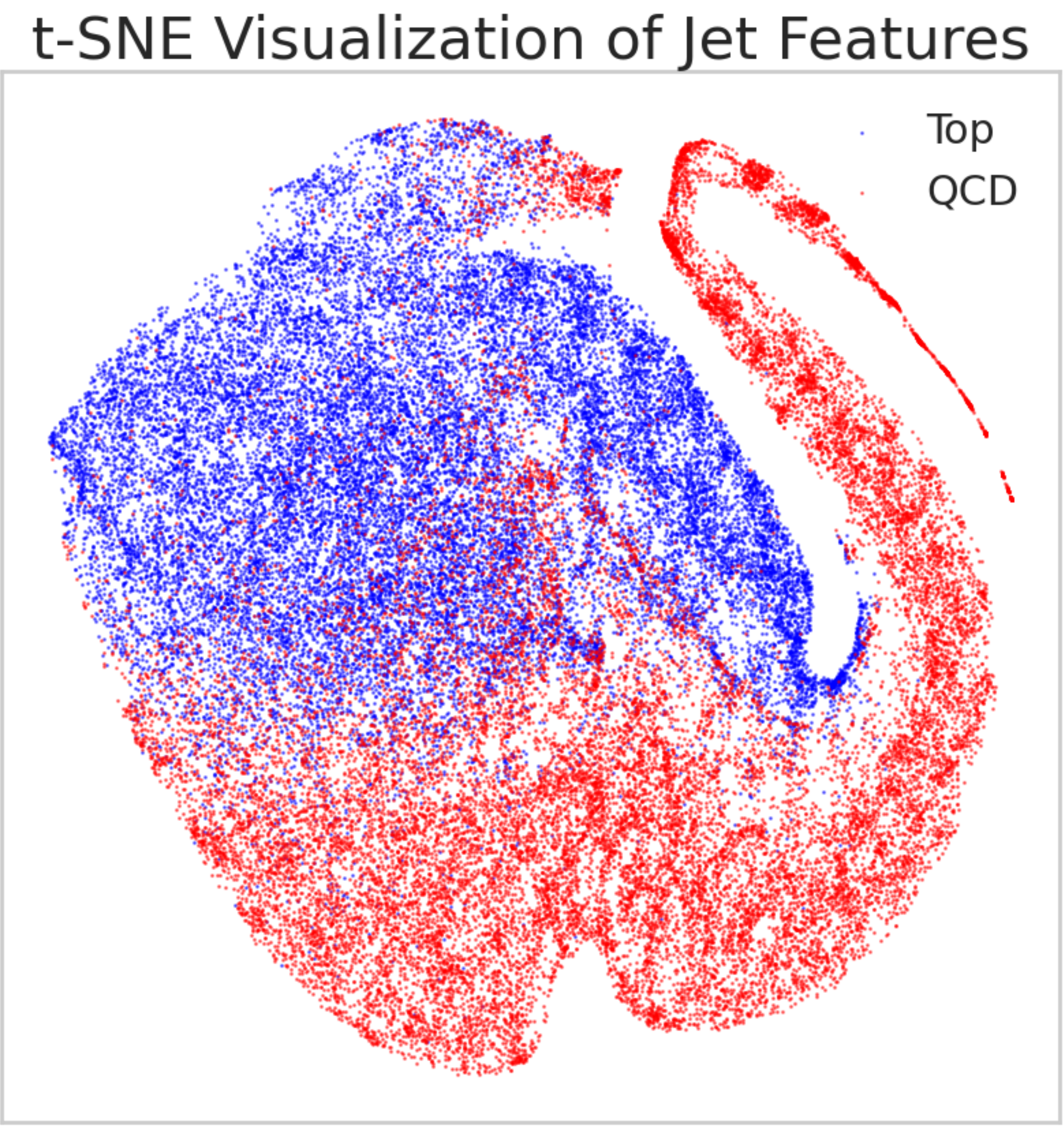
Pre-trained

Pretraining on JetClass and fine-tuning on Top Tagging

The pre-trained model shows a much clearer separation between signal and background



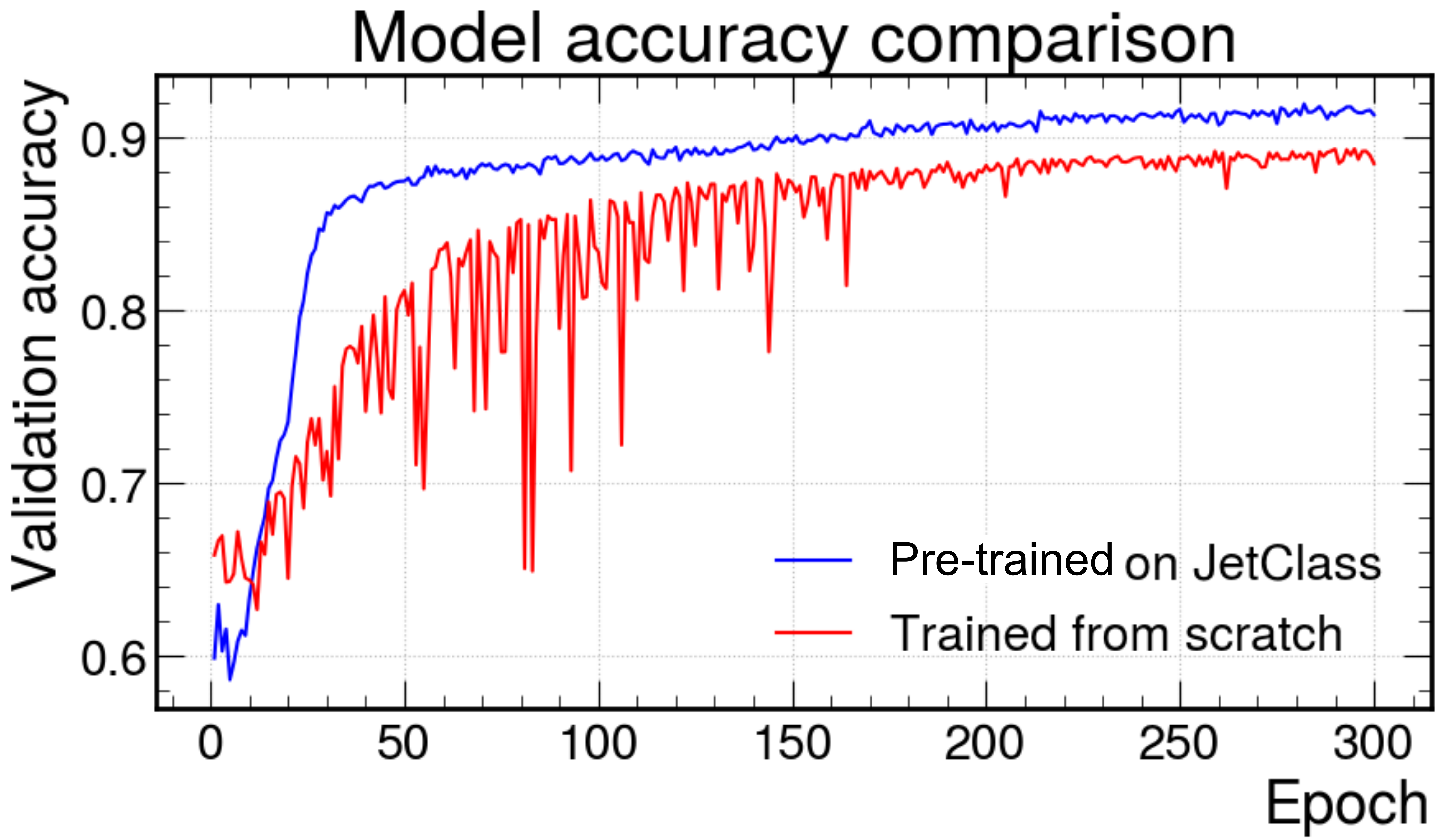
Trained from scratch



Pre-trained

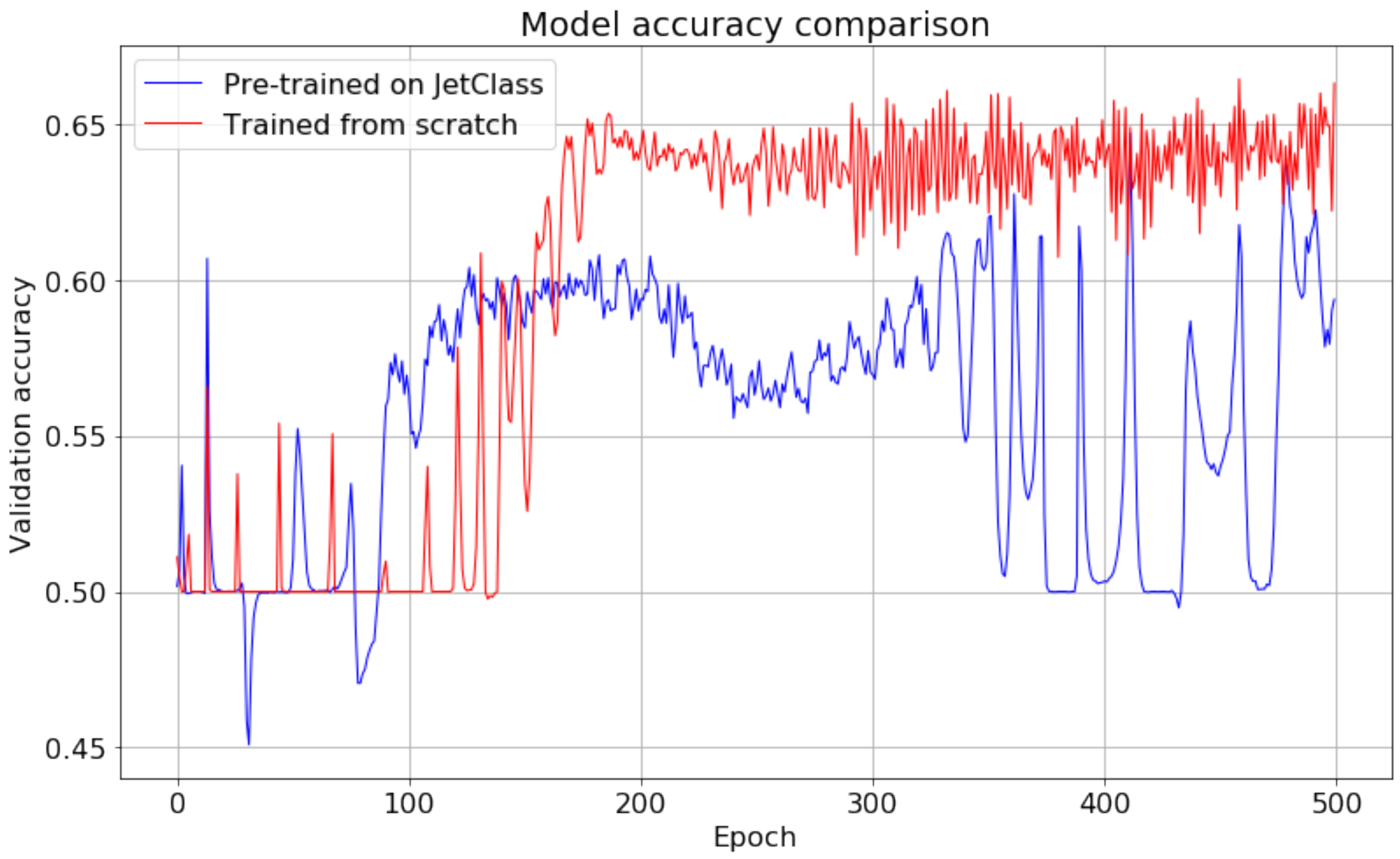
Pretraining on JetClass and fine-tuning on Top Tagging

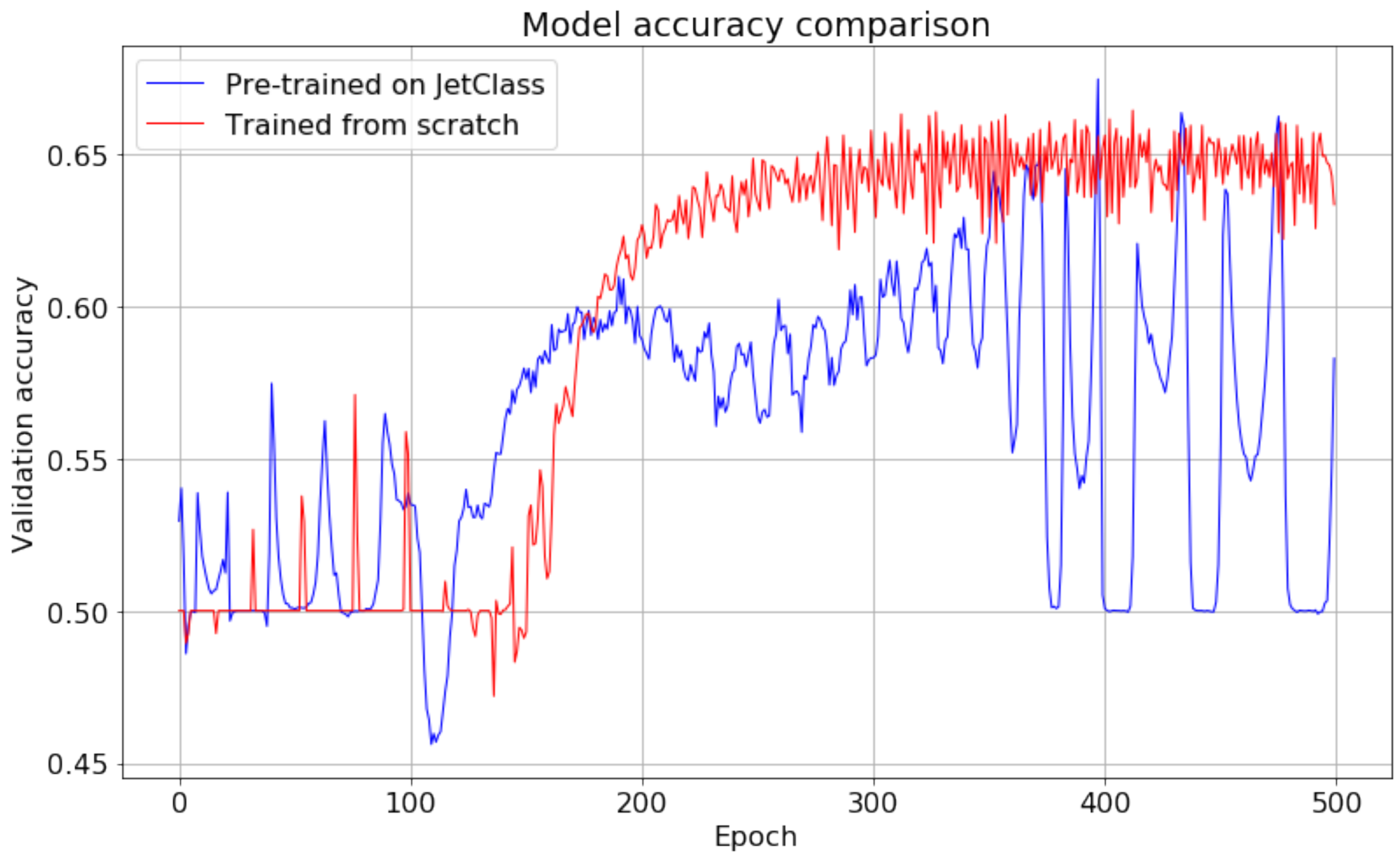
Despite limited data, the pre-trained model achieves higher accuracy and converges faster



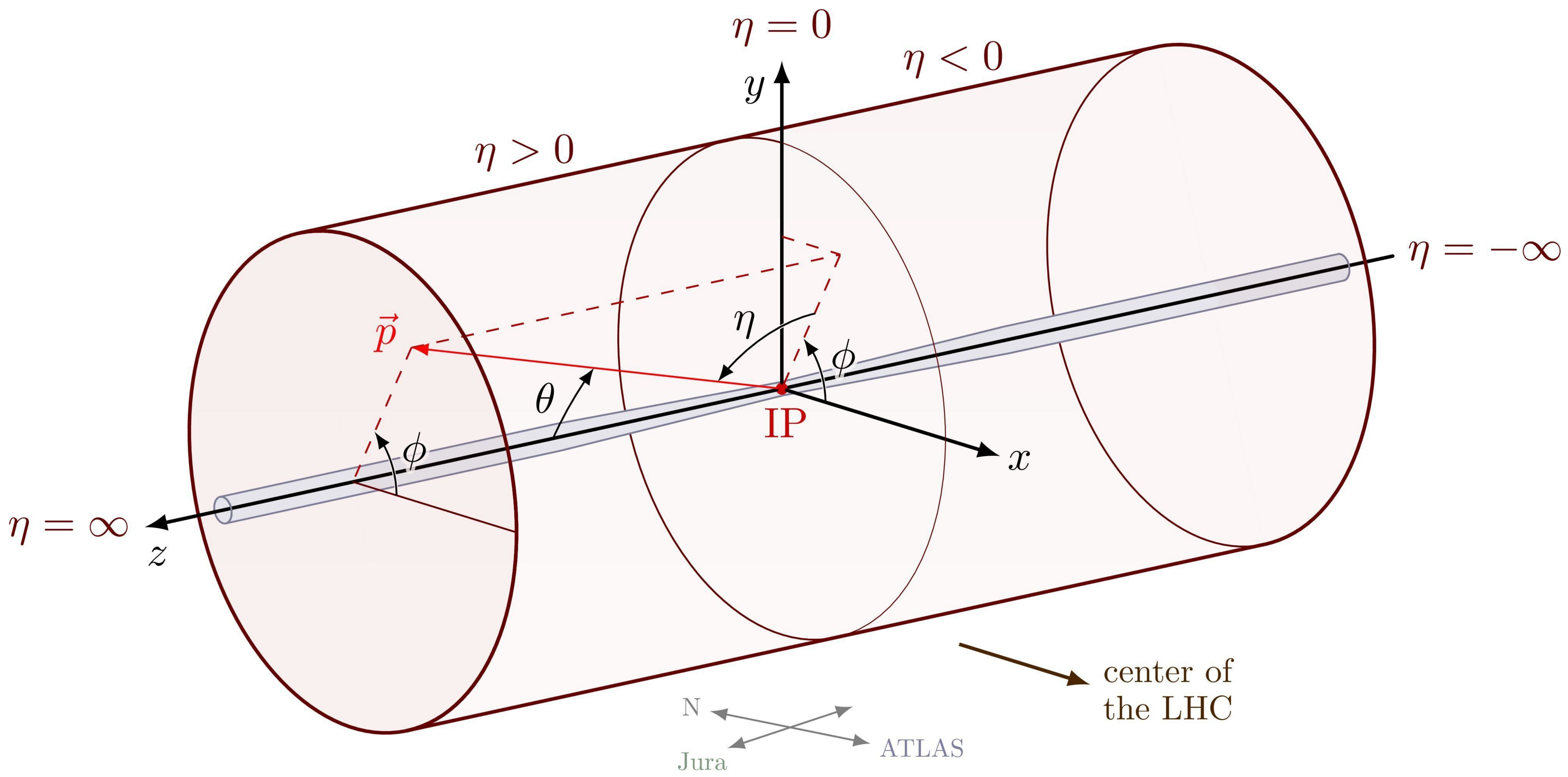
- A linear layer was added to the encoder for fine-tuning.
- Blue curve was pre-trained on 1% of the JetClass dataset (1 Million jets) with SimCLR
- Red curve was trained from scratch
- Both models share the same hyperparameters
- Both models are trained with 100k jets (1/12 of the Top Tagging Dataset)

Accuracies of two trials trained with 1000 labeled samples





The CMS detector coordinate system



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

https://tikz.net/axis3d_cms/

Details of the Top Tagging Dataset

The top signal and mixed quark-gluon background jets are produced with using Pythia8 [25] with its default tune for a center-of-mass energy of 14 TeV and ignoring multiple interactions and pile-up. For a simplified detector simulation we use Delphes [26] with the default ATLAS detector card. This accounts for the curved trajectory of the charged particles, assuming a magnetic field of 2 T and a radius of 1.15 m as well as how the tracking efficiency and momentum smearing changes with η. The fat jet is then defined through the anti-k_T algorithm [27] in FastJet [28] with $R = 0.8$. We only consider the leading jet in each event and require

$$p_{T,j} = 550 \; .... \; 650 \text{ GeV} \; . \tag{1}$$

For the signal only, we further require a matched parton-level top to be within $\Delta R = 0.8$, and all top decay partons to be within $\Delta R = 0.8$ of the jet axis as well. No matching is performed for the QCD jets. We also require the jet to have $|\eta_j| < 2$. The constituents are extracted through the Delphes energy-flow algorithm, and the 4-momenta of the leading 200 constituents are stored. For jets with less than 200 constituents we simply add zero-vectors.

1902.09914

Details of the JetClass Dataset

Simulation setup. Jets in this dataset are simulated with standard Monte Carlo event generators used by LHC experiments. The production and decay of the top quarks and the W, Z and Higgs bosons are generated with MADGRAPH5_aMC@NLO (Alwall et al., 2014). We use PYTHIA (Sjöstrand et al., 2015) to evolve the produced particles, i.e., performing parton showering and hadronization, and produce the final outgoing particles[1]. To be close to realistic jets reconstructed at the ATLAS or CMS experiment, detector effects are simulated with DELPHES (de Favereau et al., 2014) using the CMS detector configuration provided in DELPHES. In addition, the impact parameters of electrically charged particles are smeared to match the resolution of the CMS tracking detector (CMS Collaboration, 2014). Jets are clustered from DELPHES E-Flow objects with the anti-k_{T} algorithm (Cacciari et al., 2008; 2012) using a distance parameter $R = 0.8$. Only jets with transverse momentum in 500–1000 GeV and pseudorapidity $|\eta| < 2$ are considered. For signal jets, only the "high-quality" ones that fully contain the decay products of initial particles are included[2].

2202.03772

Training on Top Tagging

Are the features correlated?