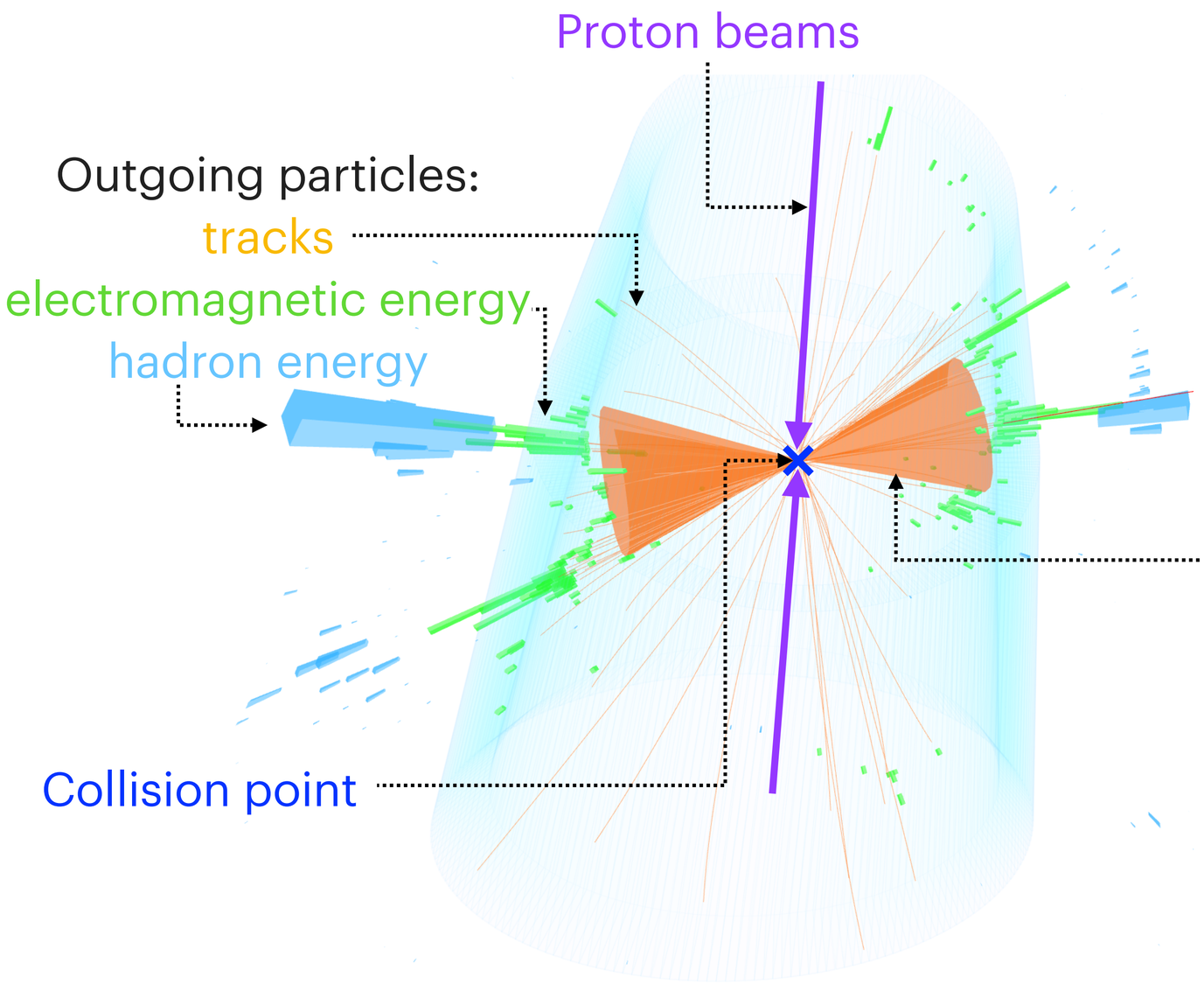
Self-Supervised Learning (SSL) for Jet Tagging

Zihan Zhao (presenter), Farouk Mokhtar, Raghav Kansal, Billy Li, Javier Duarte

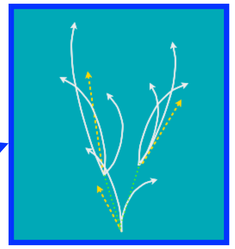
ACAT Workshop 2024
Mar 11

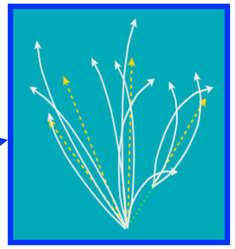
LHC and Jet Tagging

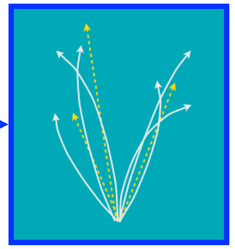


Collision event

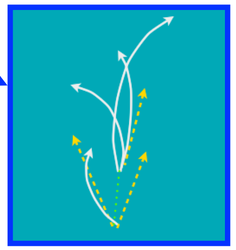
Jet


Higgs boson?


Top quark?


W or Z boson?


Gluon?


Bottom quark?

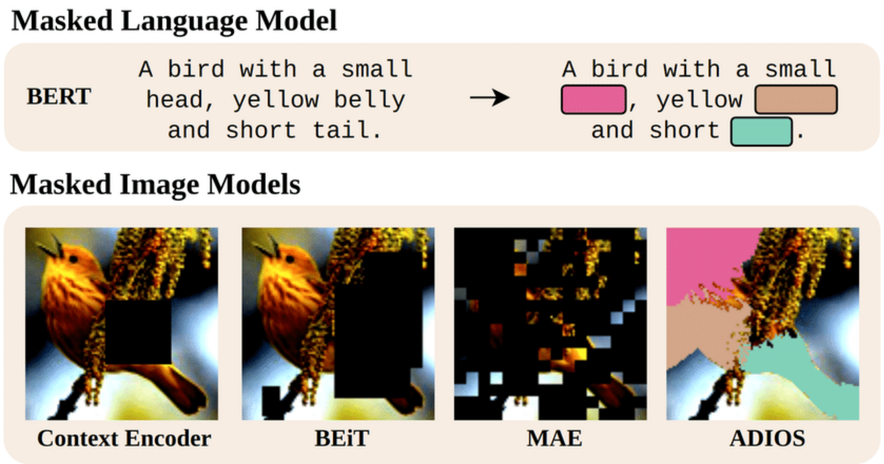
Jet tagging

1709.04464

Intro to SSL strategies

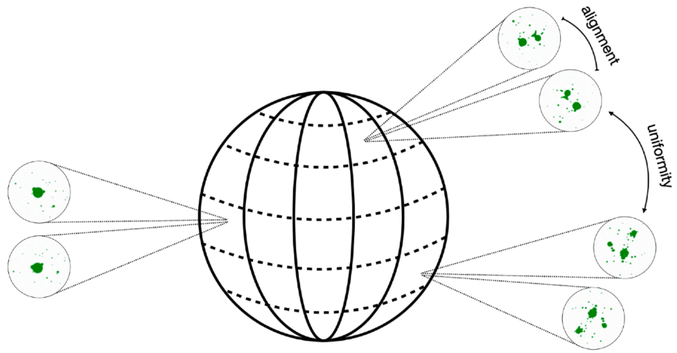
To learn useful features from the data itself without using labels

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)



Masking

2201.13100

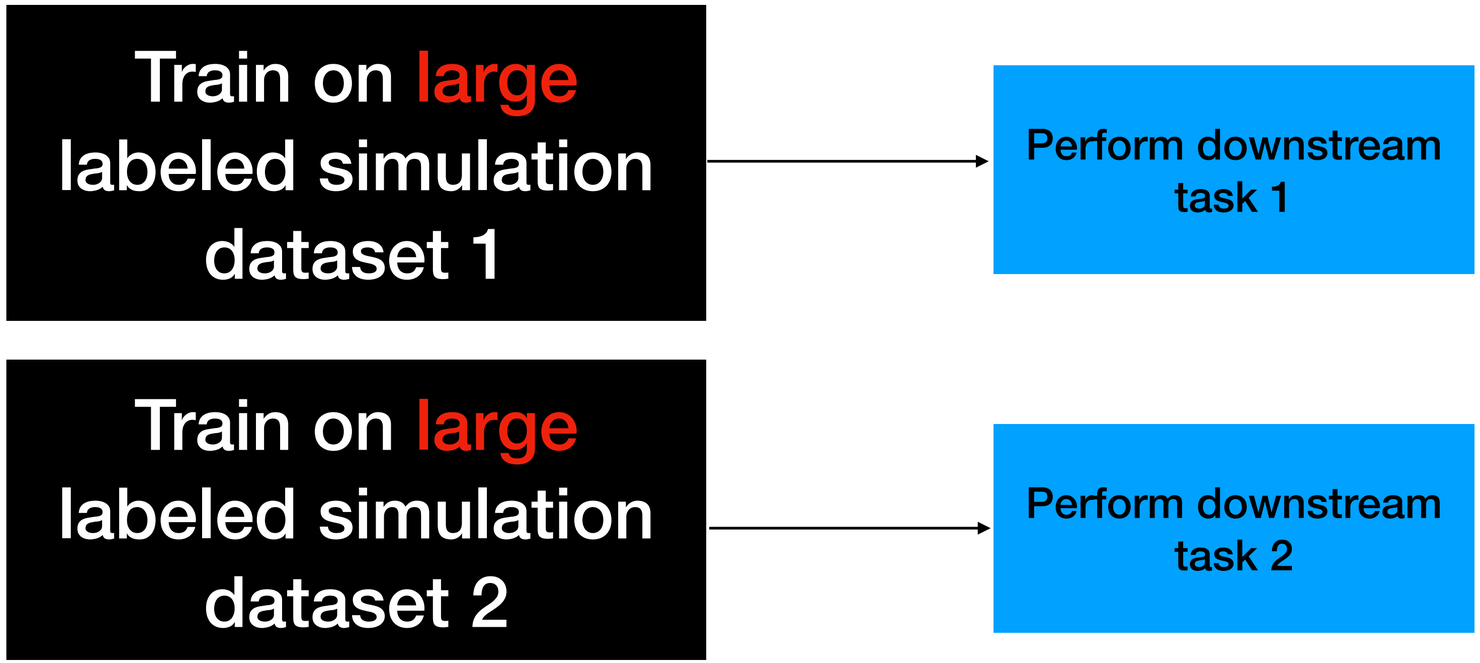


Contrastive Learning

2108.04253

Goals of the Project

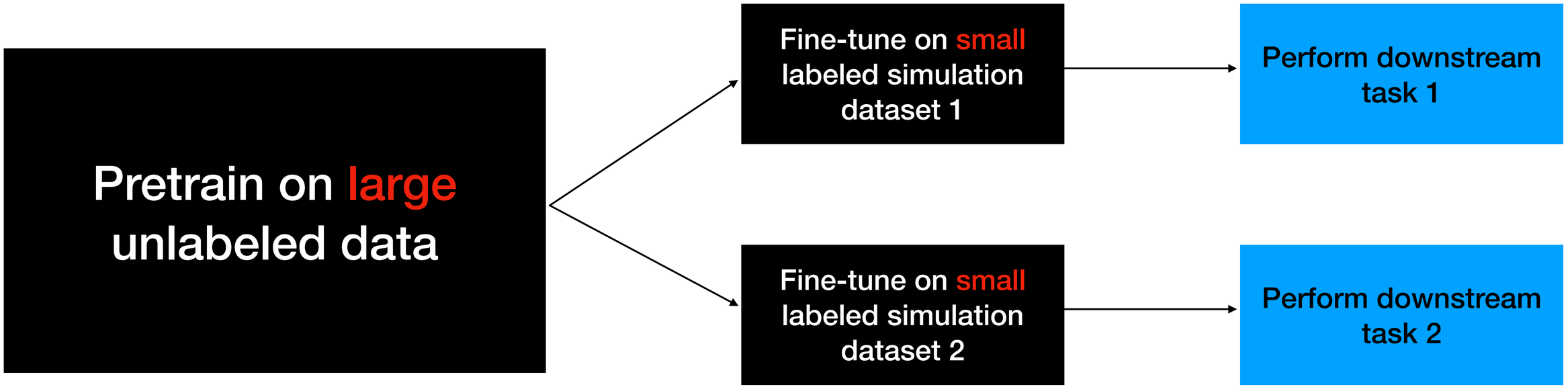
- To show that we can leverage SSL to learn powerful, generic, and transferable features directly from vast unlabeled data.



Current workflow using only Supervised Learning

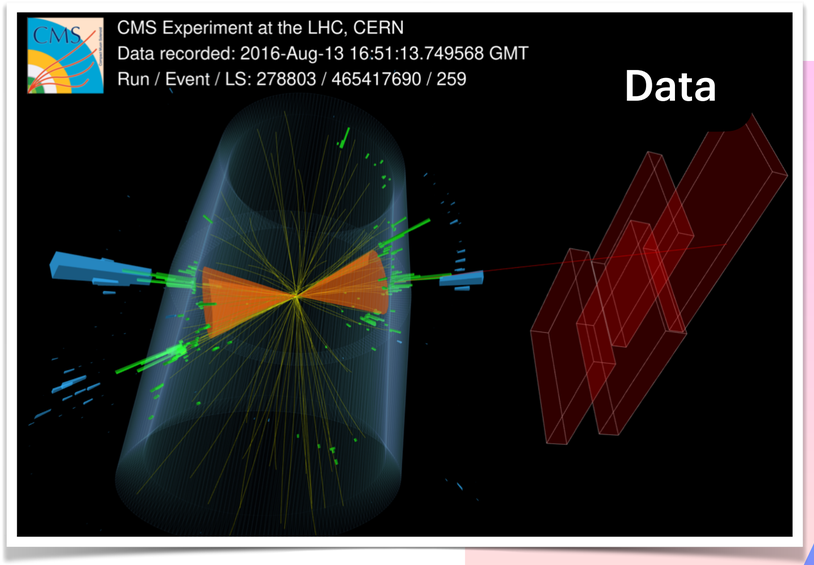
Goals of the Project

- Focus on studying the effect of scaling up the sizes of pretraining datasets on the performance of different SSL strategies.

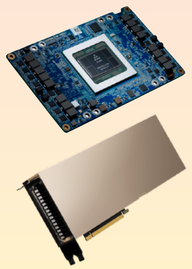


Workflow incorporating SSL

Toward Foundation Model



Training



HEP Foundation Model

Adaptation

Tasks

Jet Tagging

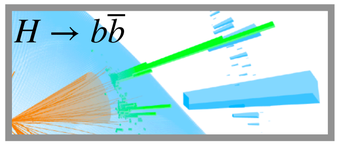
Tracking

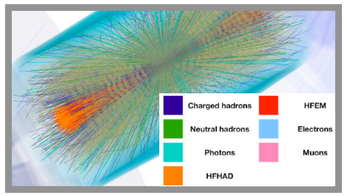
Particle-Flow Reconstruction

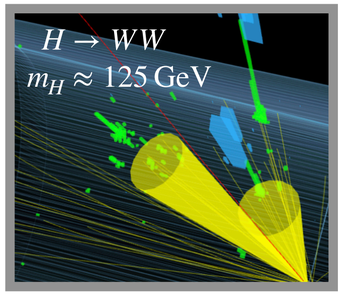
Pileup Mitigation

Jet Mass Regression

Jet Energy Correction

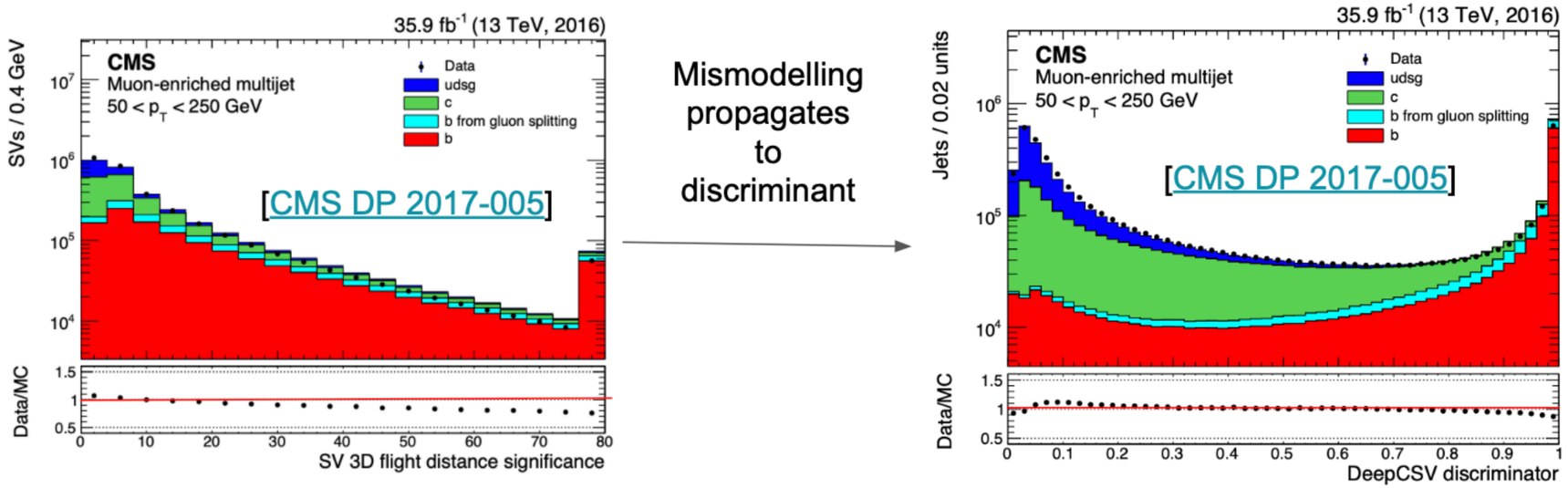






Necessity of SSL in LHC Physics

- 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

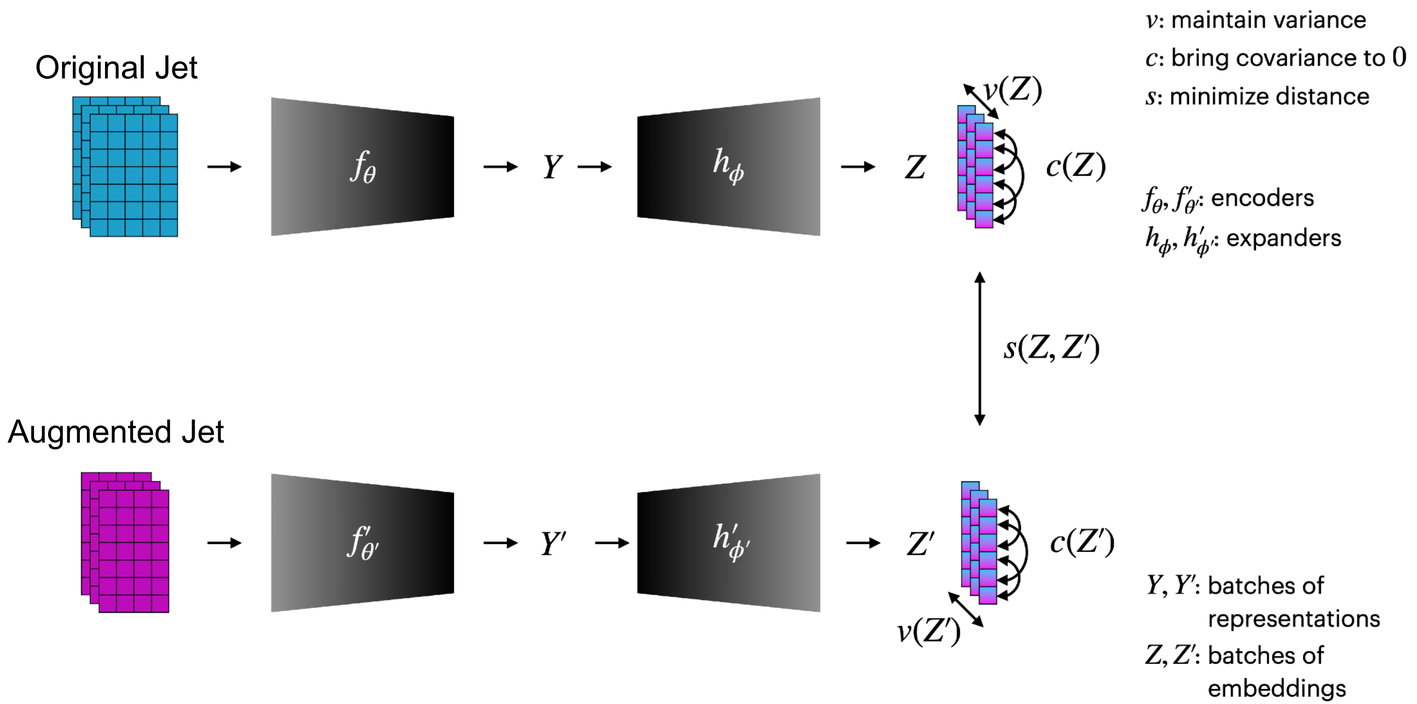
Output discriminant

Outline

- Brief intro to SSL
- Goal of the Project
- Necessity of SSL in LHC physics
- Intro to VICReg and SimCLR
- Proof of concept: Training on Top Tagging
- Transfer Learning: from JetClass to Top Tagging
- Future work

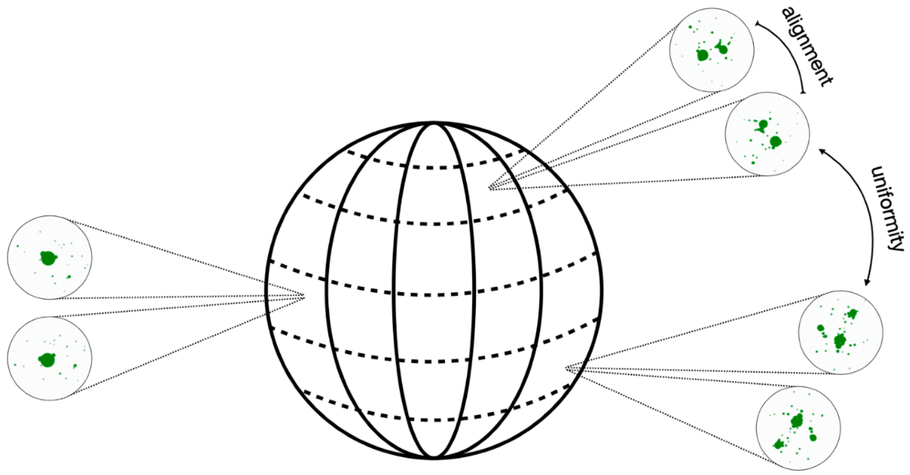
Intro to VICReg and SimCLR

general principles



VICReg: Align two views with added regularization

2105.04906



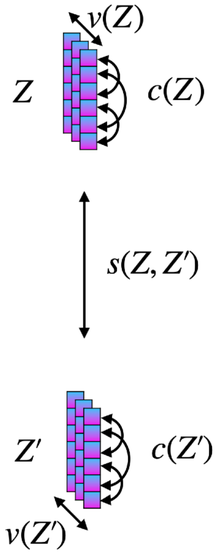
SimCLR: Contrastive learning through alignment and uniformity

2108.04253

Intro to VICReg and SimCLR

loss functions

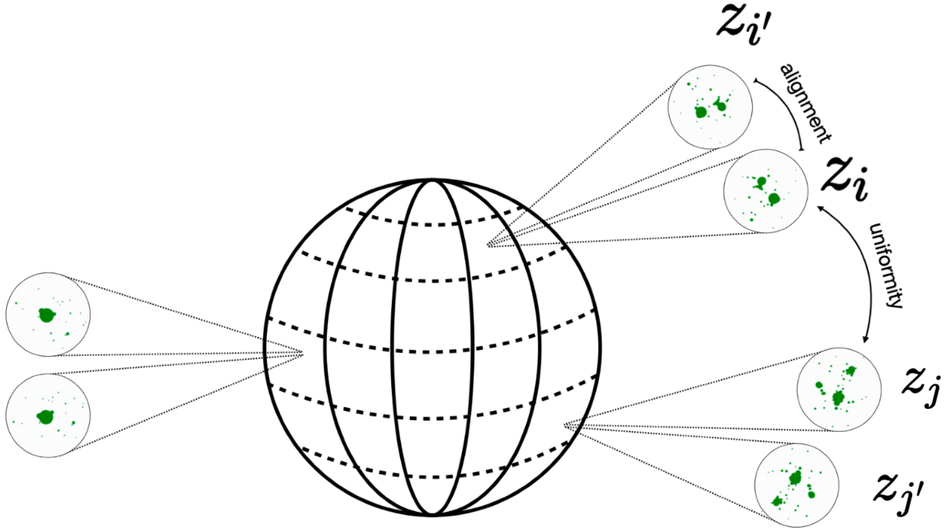
VICReg loss



$$\ell(Z, Z') = \underbrace{\lambda s(Z, Z')}_{\text{minimize distance}} + \mu[\underbrace{v(Z) + v(Z')}_{\text{maintain variance}}] + \nu[\underbrace{c(Z) + c(Z')}_{\text{bring covariance to 0}}]$$

2105.04906

SimCLR loss



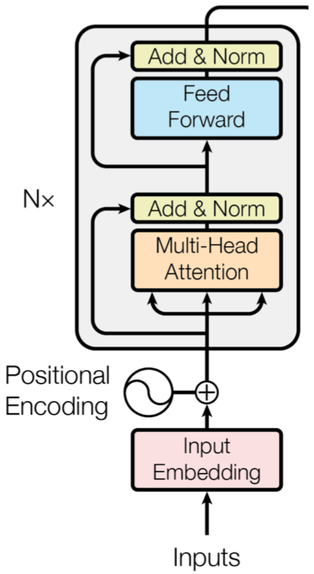
$$\mathcal{L}_i = -\log \frac{e^{s(z_i, z_i')/\tau}}{\sum_{j \neq i \in \text{batch}} \left[e^{s(z_i, z_j)/\tau} + e^{s(z_i, z_j')/\tau} \right]}$$

2108.04253

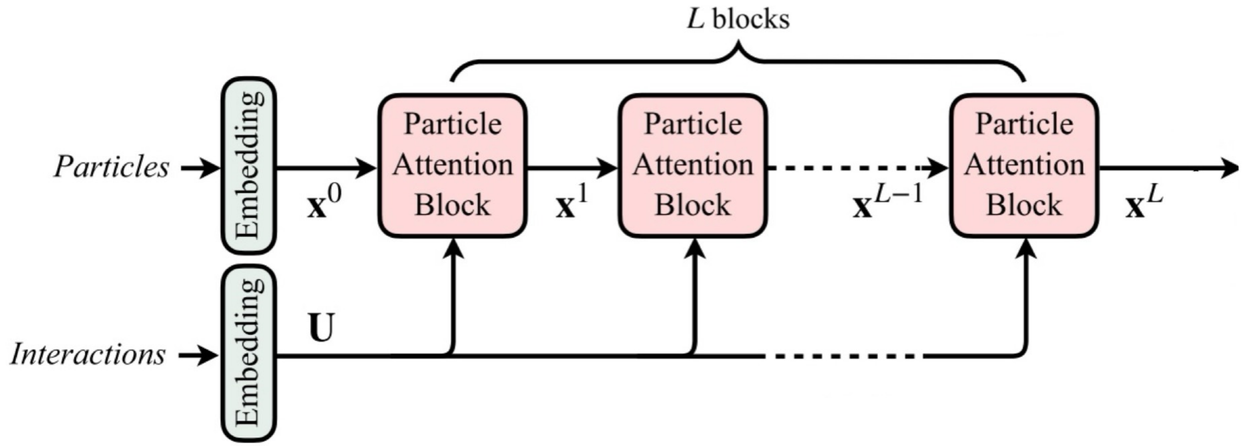
Intro to VICReg and SimCLR

Model Architecture for encoder

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



hyper-parameter	value
model (embedding) dimension	1000
feed-forward hidden dimension	1000
output dimension	1000
# self-attention heads	4
# transformer layers (N)	4
# layers	2
dropout rate	0.1



Transformer Encoder

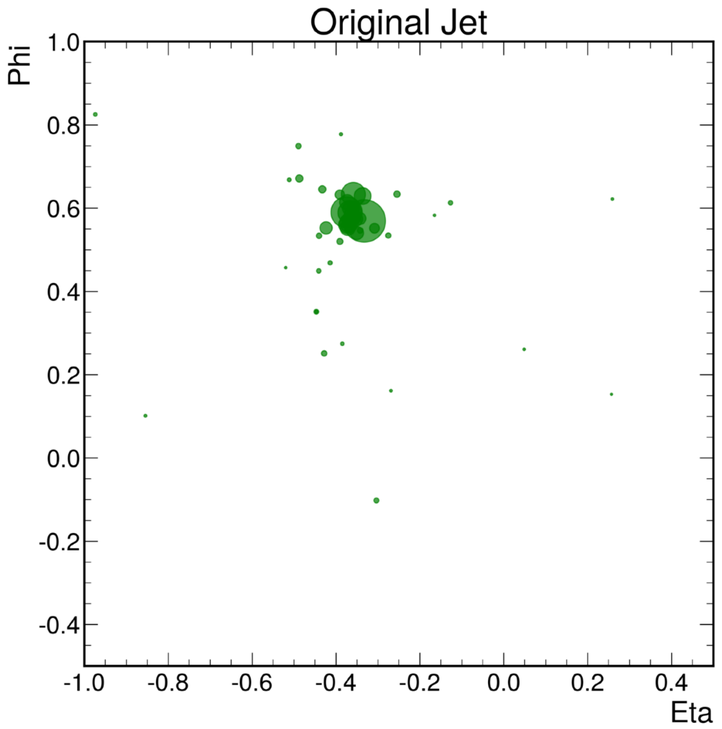
1706.03762

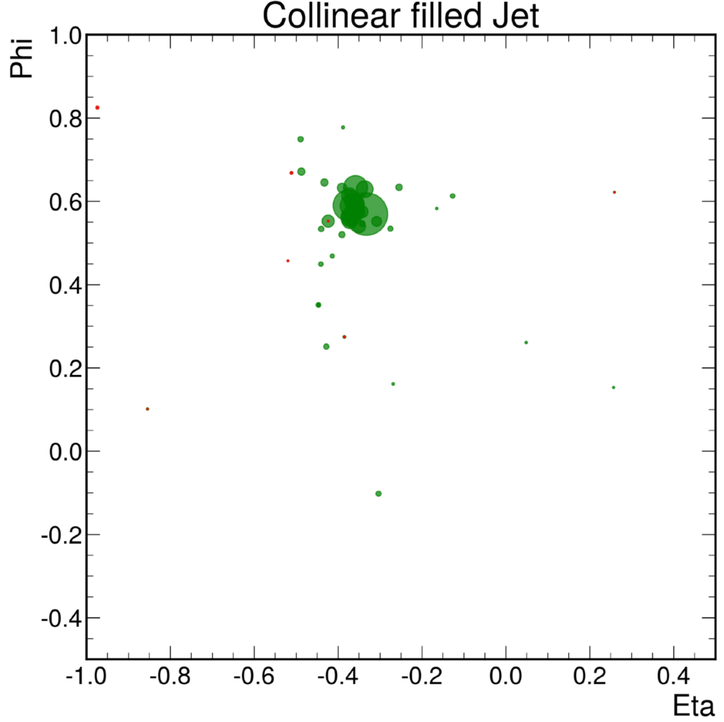
Particle Transformer

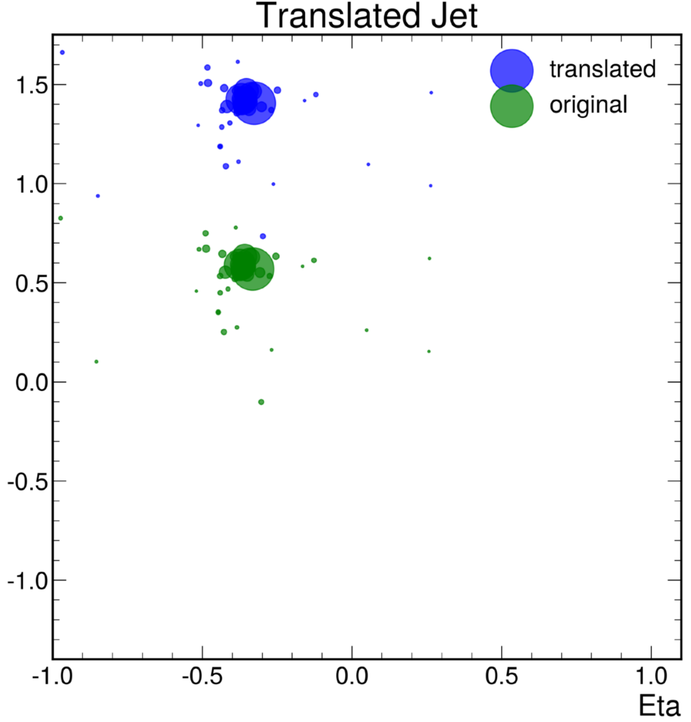
2202.03772

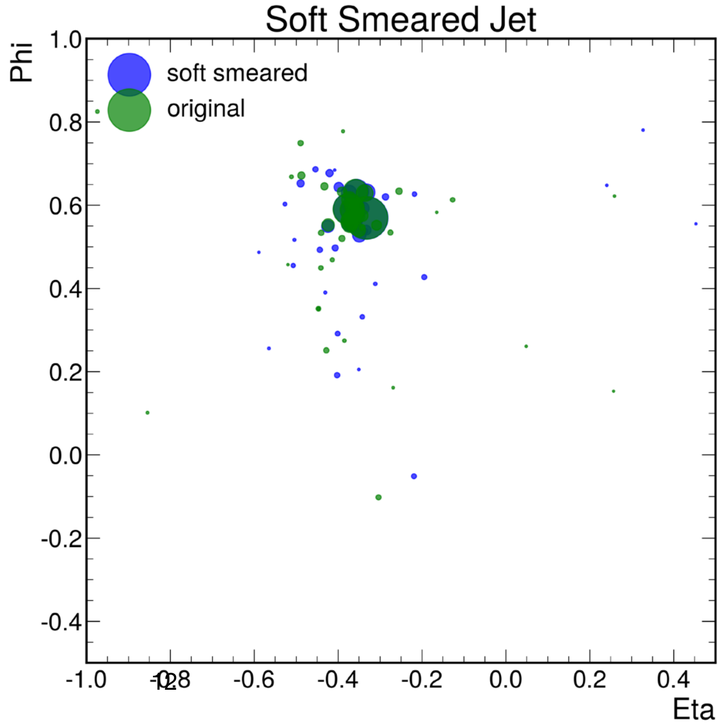
Augmentations

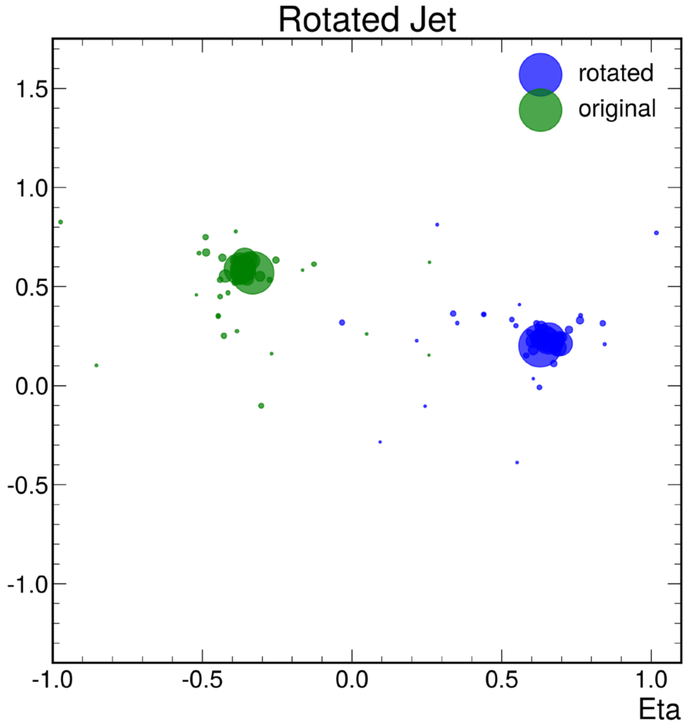
2108.04253





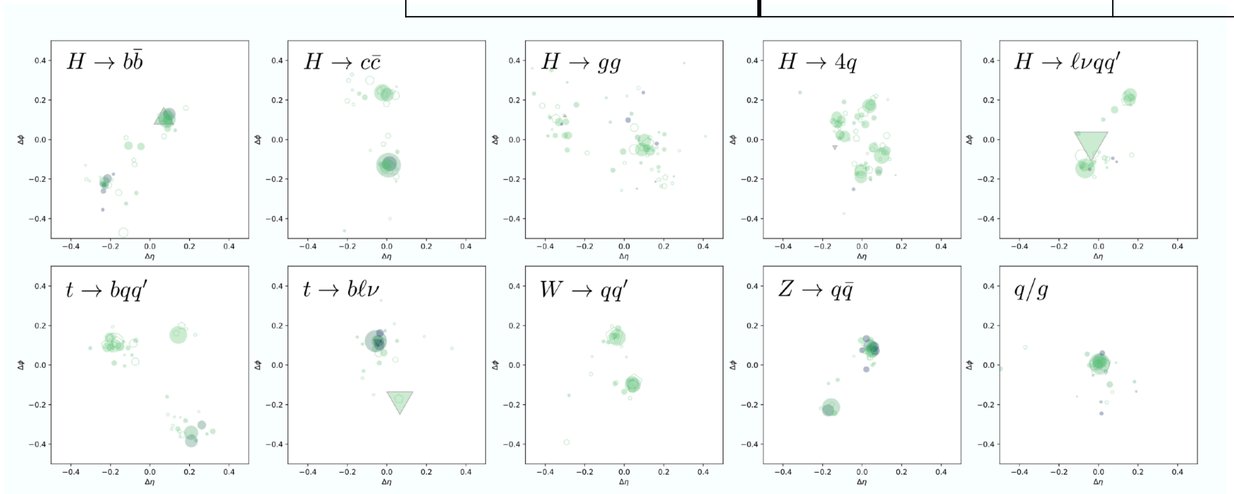






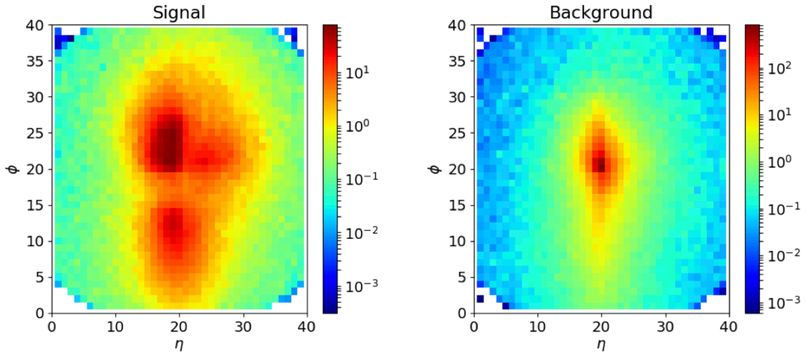
Datasets

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



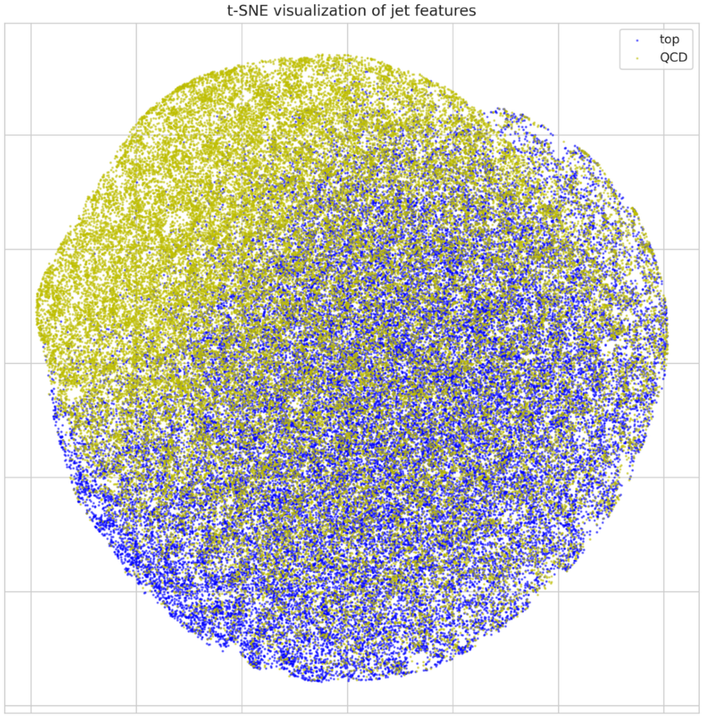
Top Tagging Dataset

1902.09914

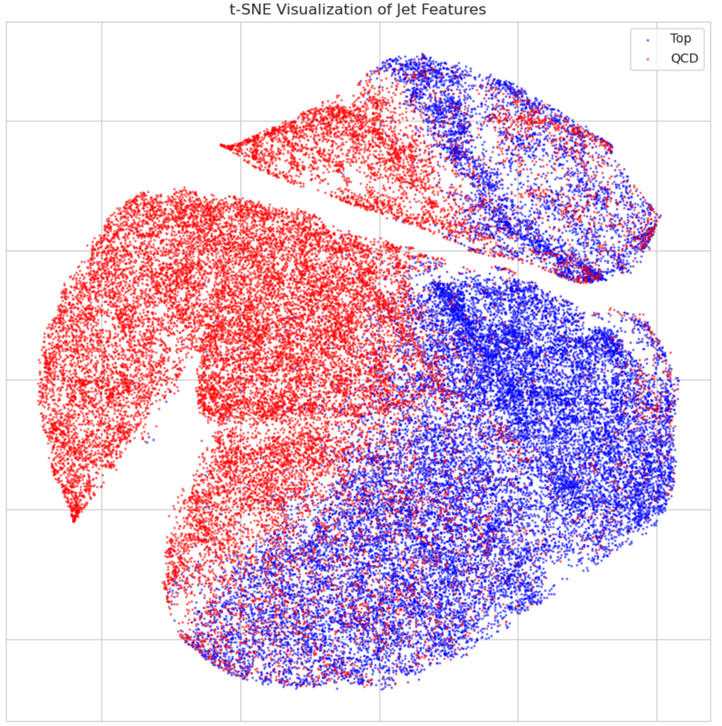
Training on Top Tagging

Comparison between VICReg and SimCLR

- SimCLR, with its clearer separation of features, outperforms VICReg in top quark jet tagging, and thus will be the main focus of our continued discussion.



VICReg learned features



SimCLR learned features

Training on Top Tagging

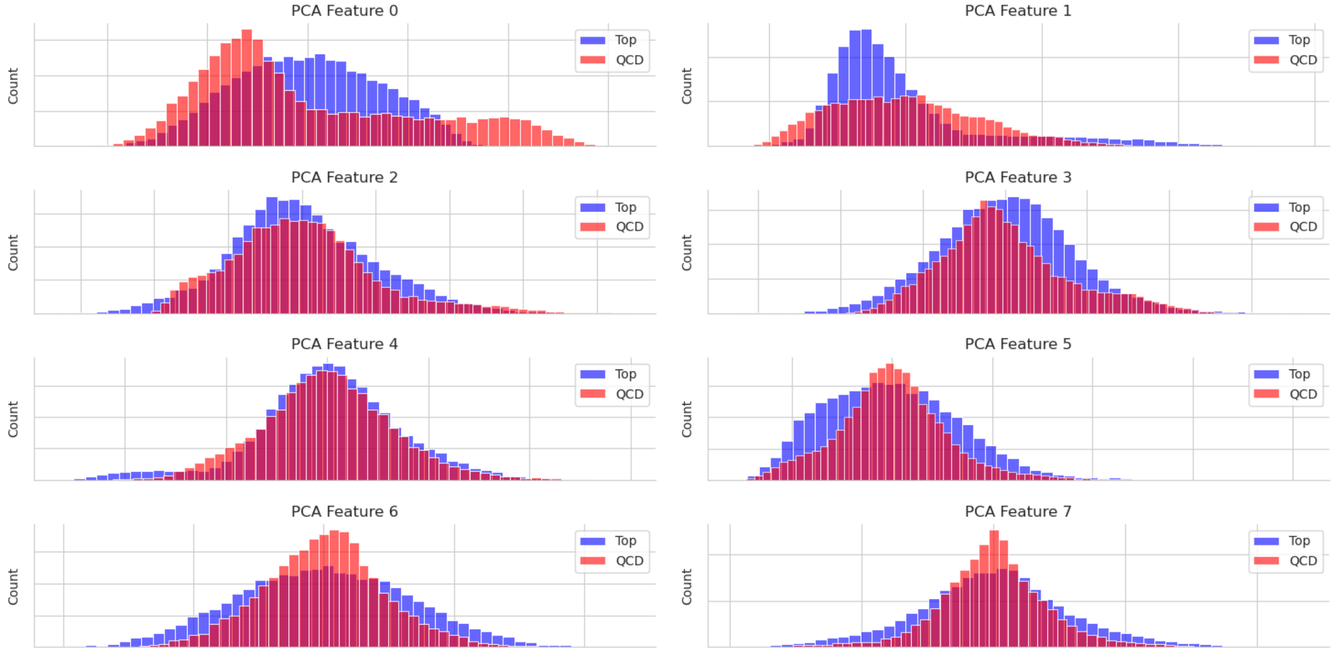
Comparison between VICReg and SimCLR

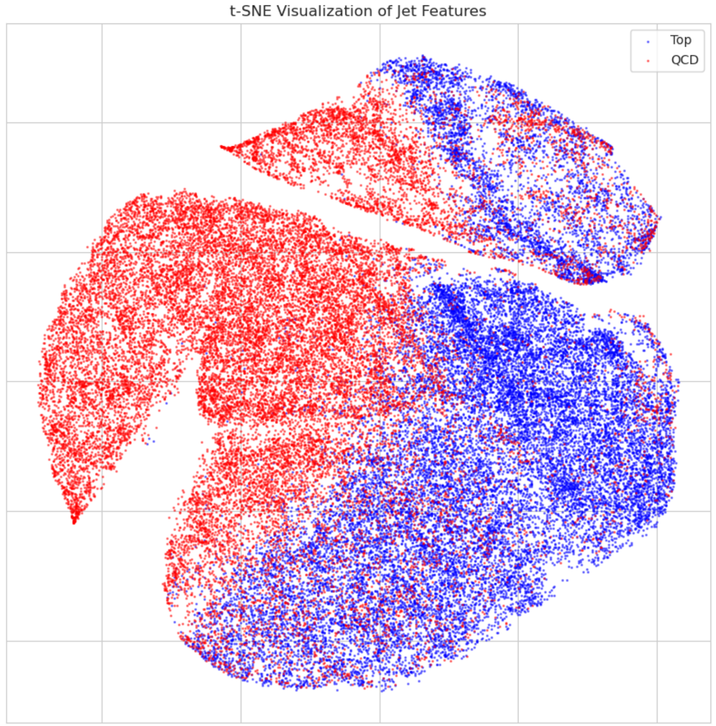
- SimCLR, with its clearer separation of features, outperforms VICReg in top quark jet tagging, and thus will be the main focus of our continued discussion.
- Potential contributing factors:
  - No explicit use of negative pairs (both pro and con)
  - Loss function has too many hyper parameters: hard to tune

Training on Top Tagging

Are the features distinguishing?

- As a proof of concept, we want to show that the model can learn some useful features that can distinguish between signal and background.

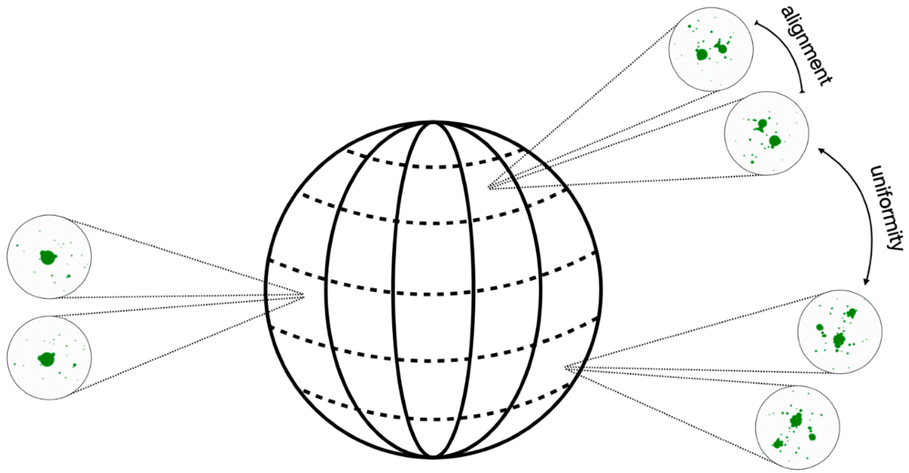




Training on Top Tagging

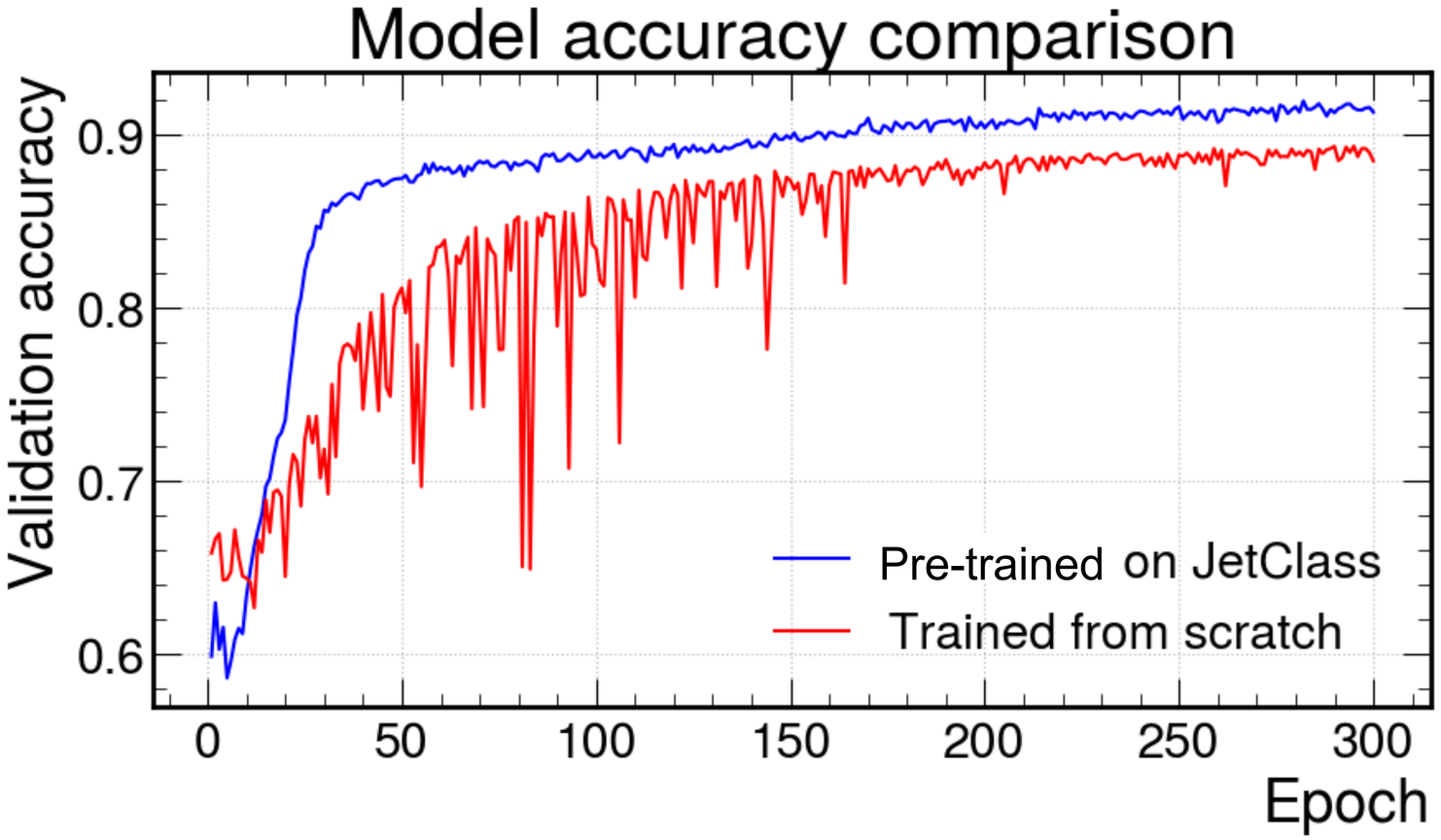
The model has learnt invariance to augmentations

Average Euclidean Distance between representations	batch 1 original	batch 1 augmented	batch 2 original	batch 2 augmented
batch 1 original	0	—	—	—
batch 1 augmented	2.10	0	—	—
batch 2 original	13.92	13.90	0	—
batch 2 augmented	13.96	13.97	2.18	0



Pretraining on JetClass and fine-tuning on Top Tagging

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



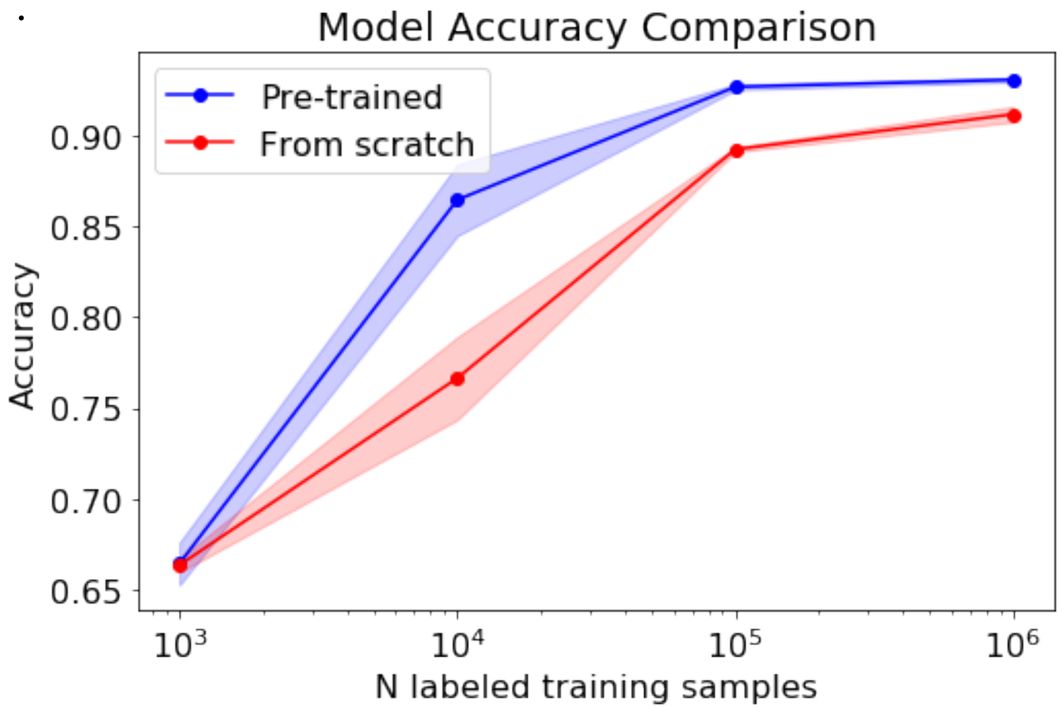
- A linear layer was added to the encoder for fine-tuning.
- Blue curve was pre-trained on 1% of the JetClass dataset (1 Million jets) with SimCLR
- Red curve was trained from scratch
- Both models share the same hyperparameters
- Both models are trained with 100k jets (1/12 of the Top Tagging Dataset)

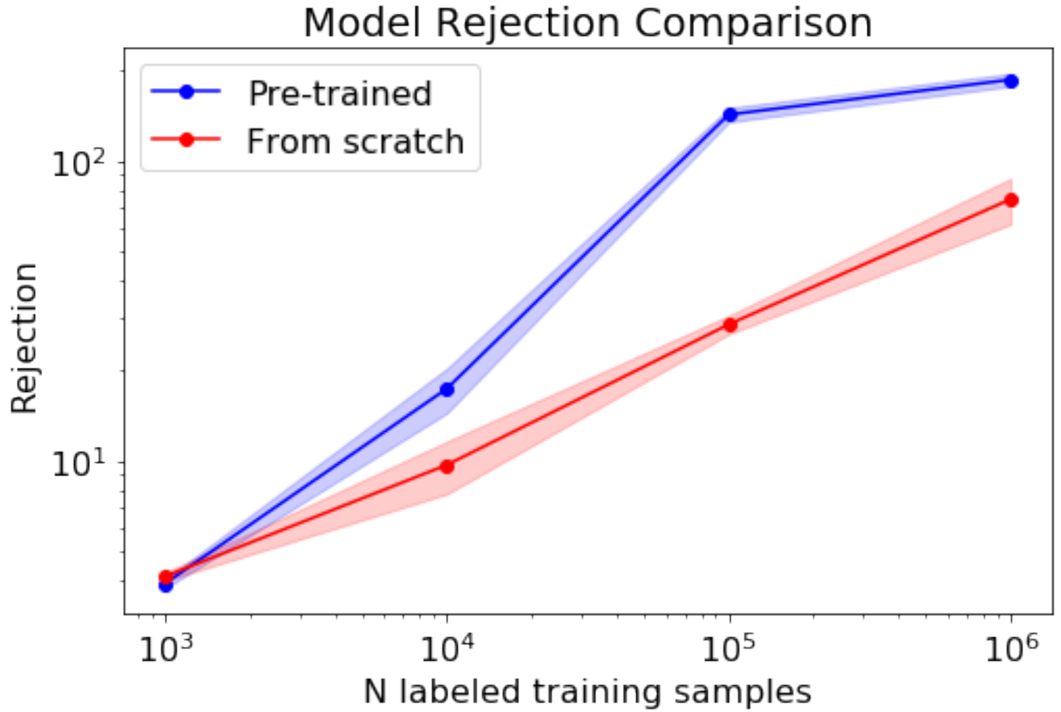
Pretraining on JetClass and fine-tuning on Top Tagging

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

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

.





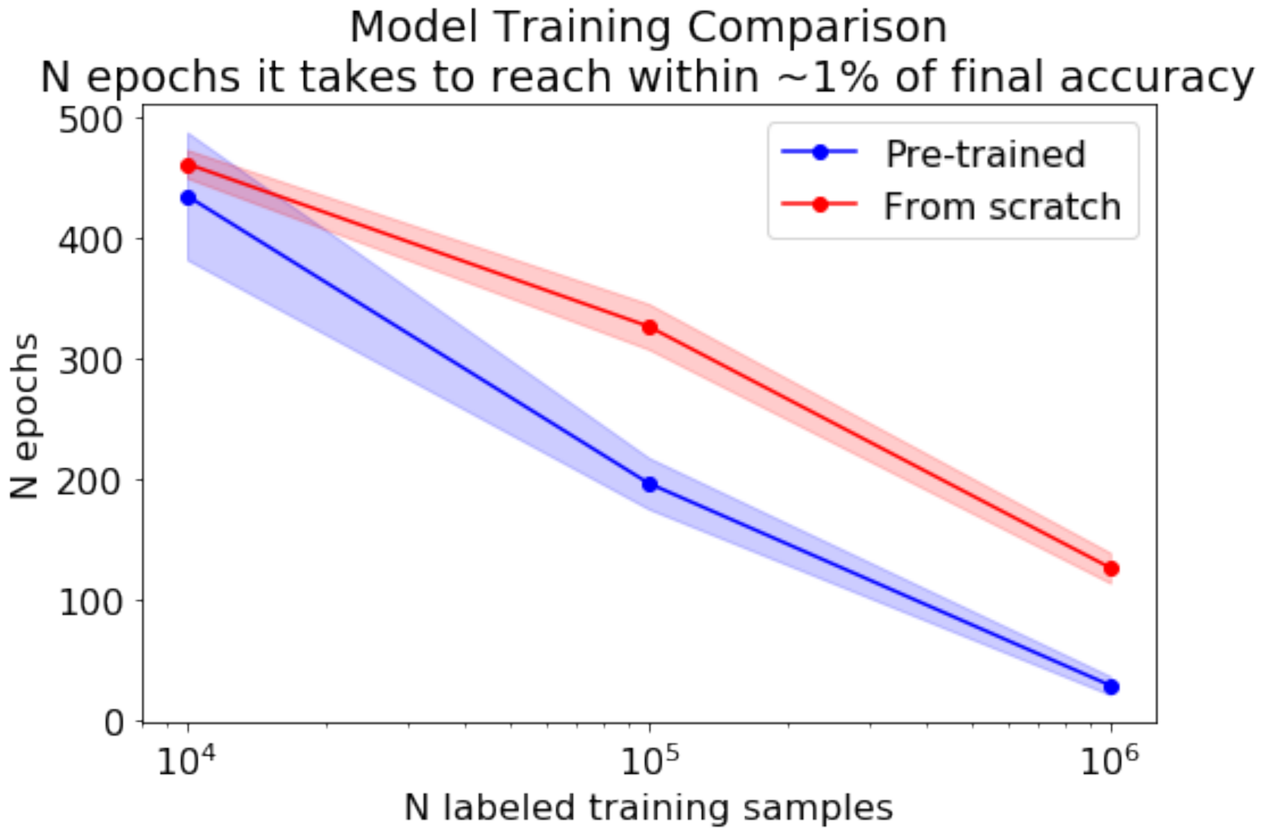
Rejection: inverse of background rejection at 50% signal efficiency

Pretraining on JetClass and fine-tuning on Top Tagging

The pre-trained model converges much faster

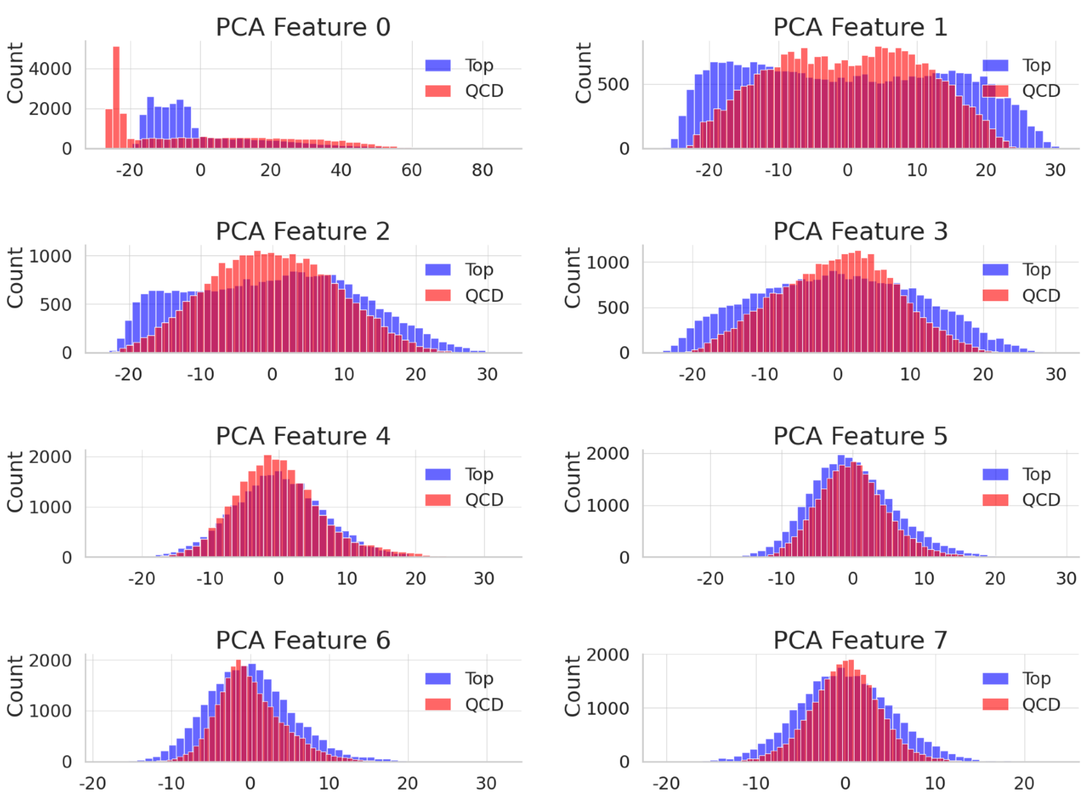
- The averages and standard deviations over 5 trainings are shown in solid lines and uncertainty bands, respectively

.

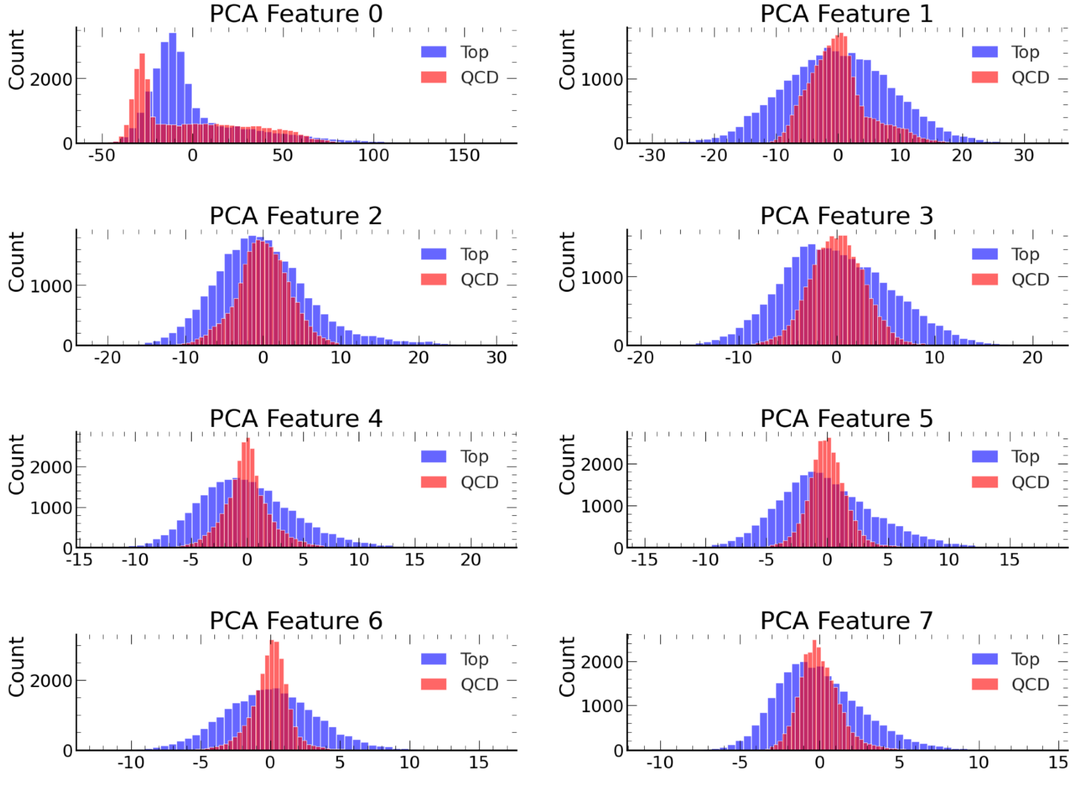


Pretraining on JetClass and fine-tuning on Top Tagging

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



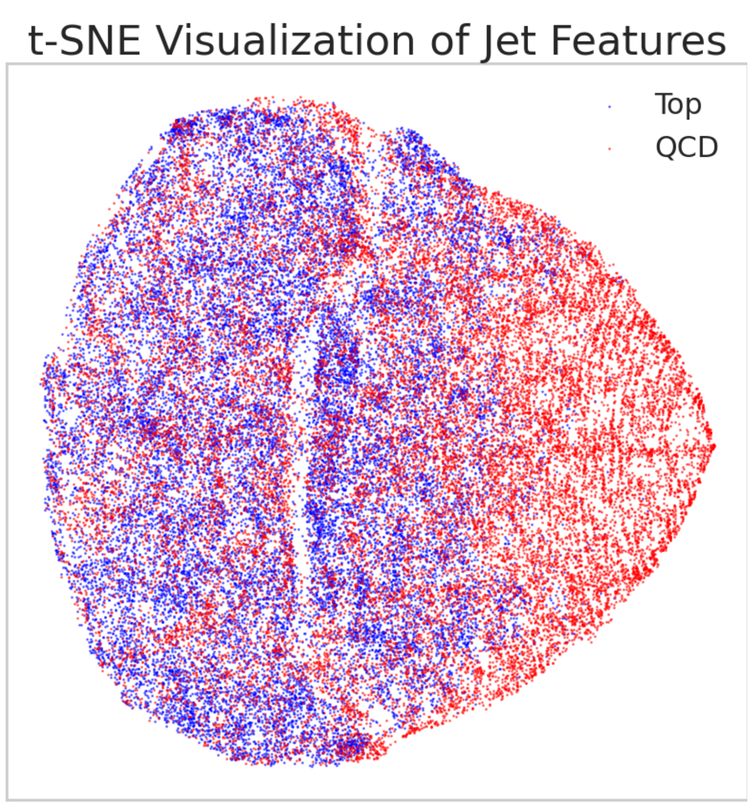
Trained from scratch



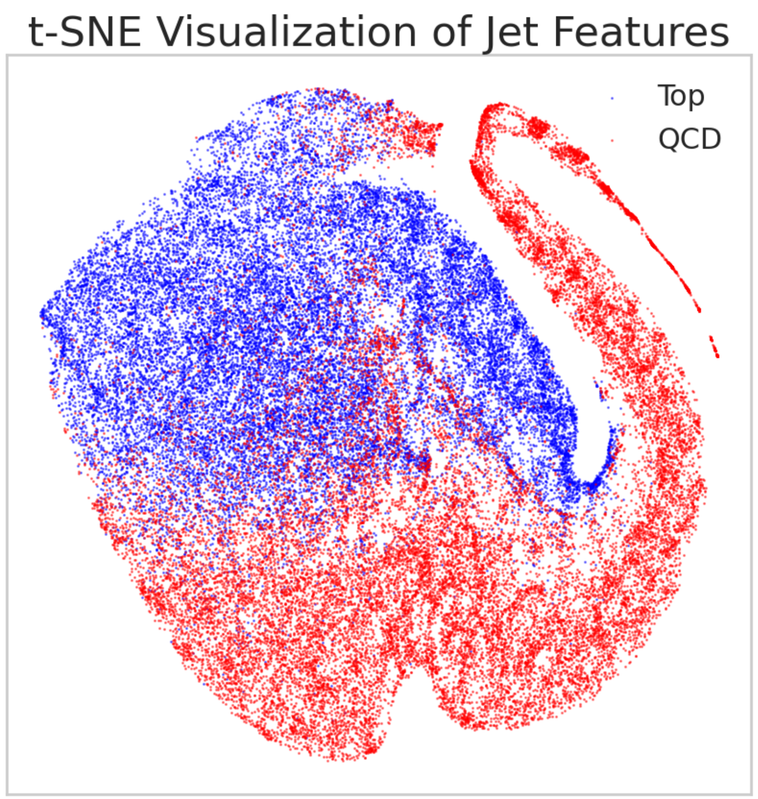
Pre-trained

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

Conclusion

- Through contrastive learning, a vanilla transformer encoder was able to learn useful representations of jets from unlabeled data.
- By pre-training on unlabeled data, the transformer encoder was able to learn the downstream task faster and with fewer labeled training samples, compared with one we trained from scratch.

Future work

- Study the scalability of dataset size in pretraining
- Study the effectiveness of more advanced architectures like the ParticleTransformer as the backbone encoder
- Explore other physically motivated augmentations
  - Pairing the two jets from dijet events
  - Using two subjets clustered with smaller radii
  - …

Support

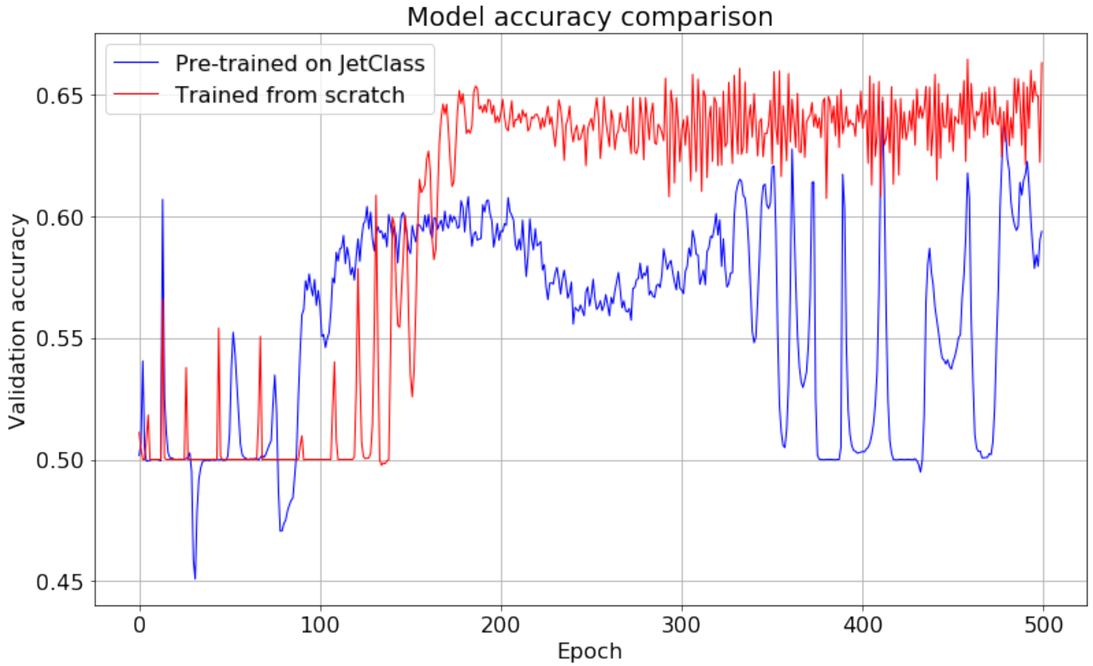
Thank you for listening!

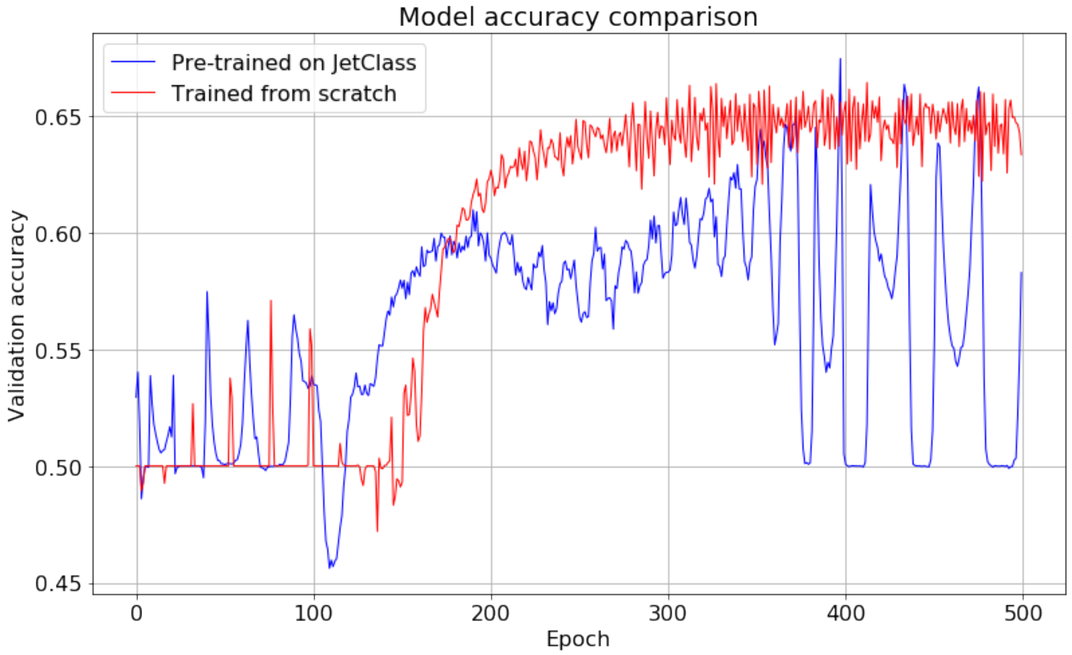
- This work is supported by the National Science Foundation under award number 2117997 (A3D3 Institute), Research Corporation For Science Advancement, and the Alfred P. Sloan Foundation

ALFRED P. SLOAN FOUNDATION

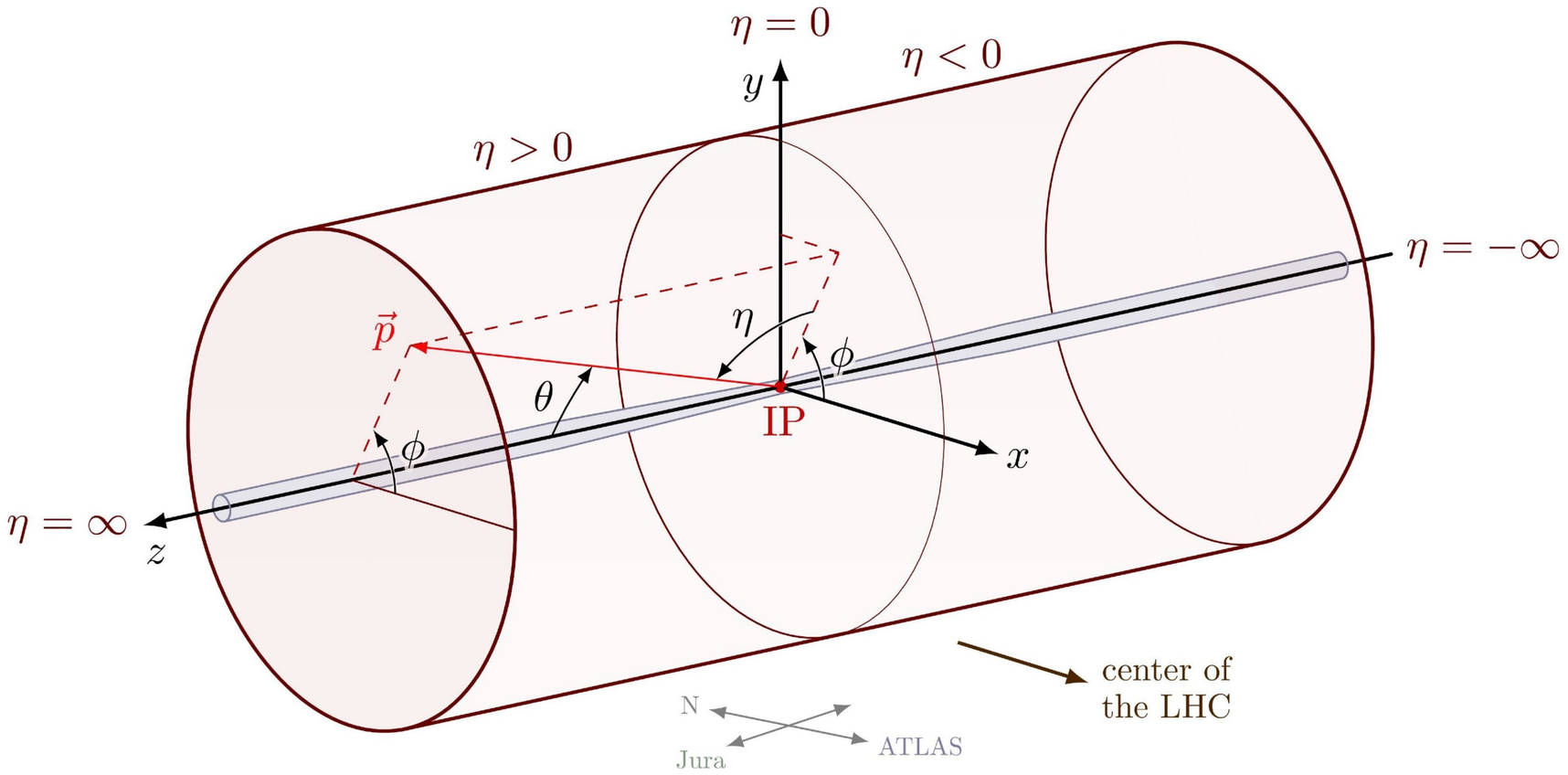
Back Up

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

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

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?