Re-simulation-based self-supervised learning (RS3L): a backbone for HEP

J. Krupa, B. Maier, M. Kagan, P. Harris, M. Pierini, N. Woodward

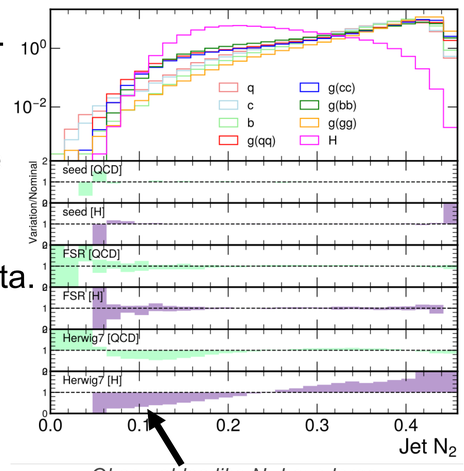
Inter-experiment Machine Learning Workshop, CERN, 1.2.2024

Karlsruher Institut für Technologie

The NSF Institute for Artificial Intelligence and Fundamental Interactions

Introduction

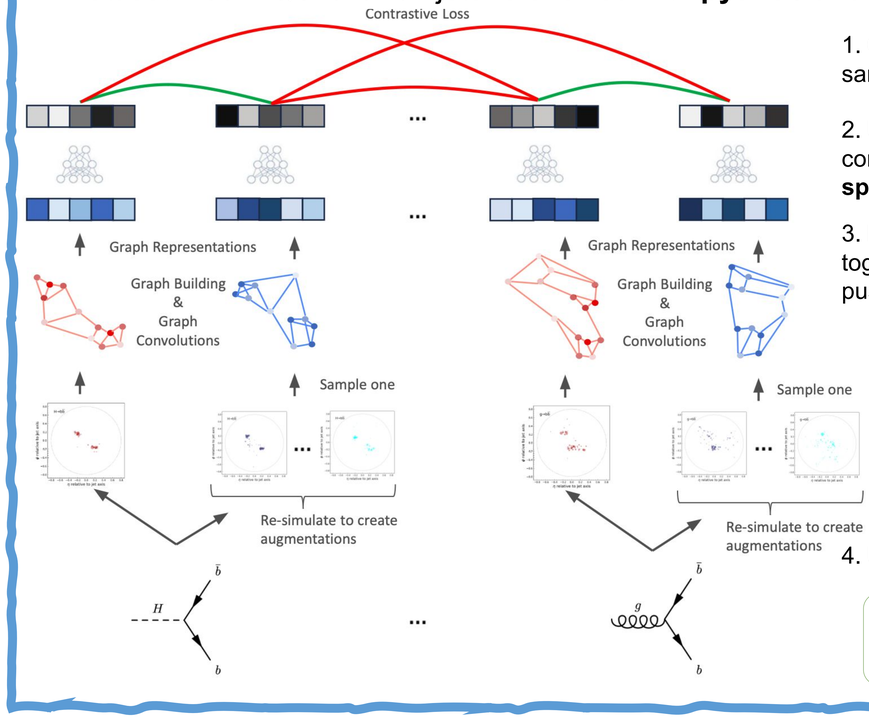
- We aim to learn a **generic** representation of jets with self-supervised learning (**SSL**), first done by [1]
- Idea: **re-simulate** an event multiple times with different simulator configurations and train a network to map these physically-motivated variations to the same point
  - Following a self-supervised method → no labels
- This is a step towards a **foundational model for HEP** data. Can our SSL model …
  1. be used in downstream tasks: classification, regression, anomaly detection, data/MC tuning, …
  2. project away uncertainties in physics modelling (e.g. mitigating uncertainties)
  3. be more efficient: give fewer training examples to get the same performance



Observables like N_2 have large variations with respect to simulator

Methods

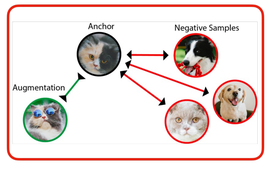
- Dataset: 100M QCD/H/W/Z jets re-showed with **pythia8** and **herwig7** tunes



1. Generate H and QCD. **Re-shower** the same events in multiple configurations

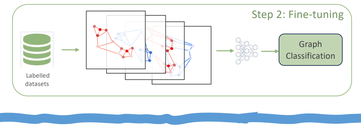
2. Use a graph NN [2] with dynamic edge convolutions to embed event into an **8D space**

3. **Build a representation** space by pulling together (anchor, augmentation) pair, pushing apart the rest using simCLR loss [3]



no labels!

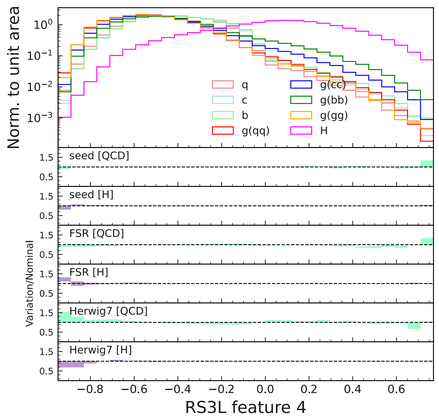
4. Use the graph for **downstream tasks**



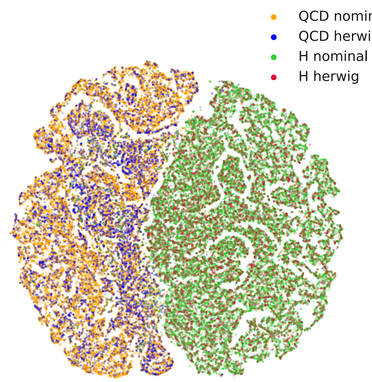
labels!

Results

1. SSL **discriminates** between Higgs and QCD + **mitigates variations** with simulator



Dimension reduction on 8 RS3L features



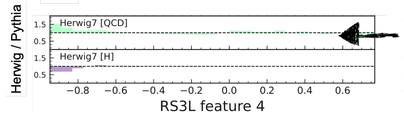
2. SSL reaches fully-supervised performance with **fewer examples** on in-domain classification

Higgs efficiency	0.3	0.5	0.7
RS3L + FT (3M, floating)	**1340**	**379**	**135**
Fully-supervised (8M)	1271	378	131

3. SSL **surpasses fully-supervised performance** on out-of-domain classification (W→qq vs QCD)

W efficiency	0.3	0.5	0.7
RS3L + FT (3M)	**1893**	**505**	**147**
Fully-supervised (3M)	1781	457	134

4. The Wasserstein distance between the classifier score as evaluated on nominal and herwig jets is reduced by RS3L → SSL can provide **more robust observables**

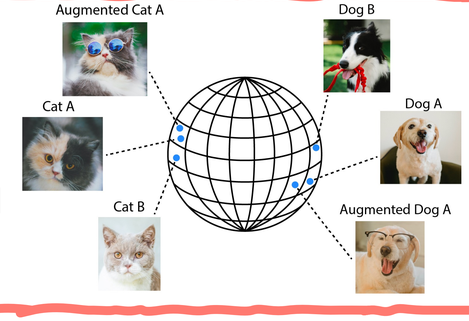


How much work does it take to transform pythia distribution into herwig?

Training setup	Herwig
Fine-tuned (3M, fixed)	7.20×10^{-3}
Fine-tuned (3M, floating)	7.80×10^{-3}
Fully-supervised (8M)	9.40×10^{-3}

Conclusions

- SSL has large potential in downstream HEP applications
  - Both in-domain and out-of-domain
- A path towards a **foundation model** for LHC physics
  - Can we train on data?



References

[1] Dillon et al, "Symmetries, Safety, and Self-Supervision", 2021, arXiv.2108.04253.
[2] Qu et al, "ParticleNet: Jet Tagging via Particle Clouds", Phys Rev D 101 056019 (2020), arXiv:1902.08570.
[3] Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations, ICML2020. arXiv:2002.05709.