Finetuning Foundation Models for Joint Analysis Optimisation

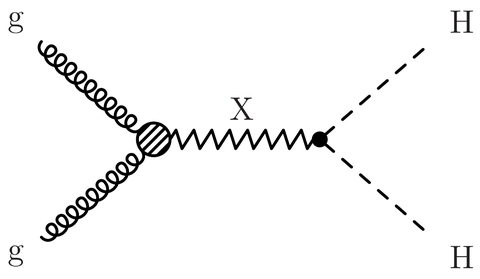
Matthias Vigl[a], Nicole Hartman[a], Lukas Heinrich[a]

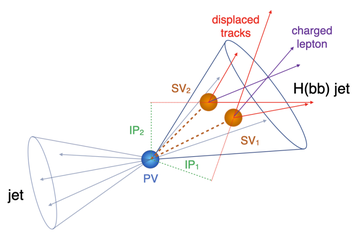
a: Technical University of Munich

arXiv:2401.13536

IML Workshop CERN, Switzerland, 29.01.2024 – 02.02.2024

A toy end-to-end Analysis

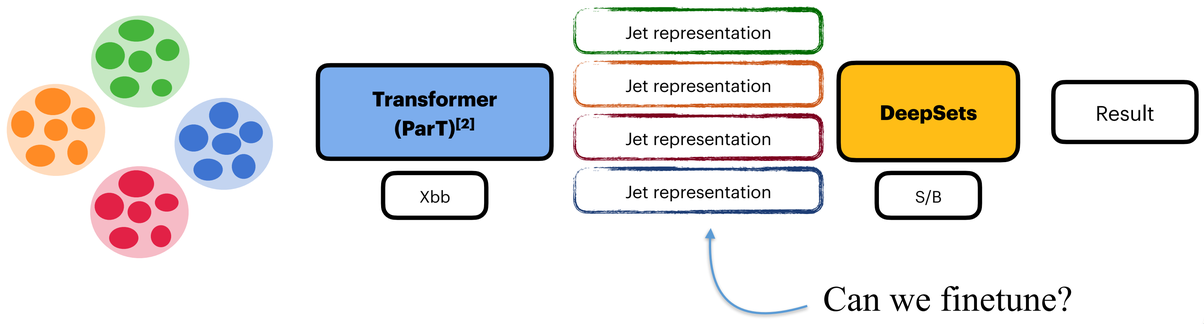




Multi-jet final state with $X \to HH \to b\bar{b}b\bar{b}$[1]. Typical strategy:

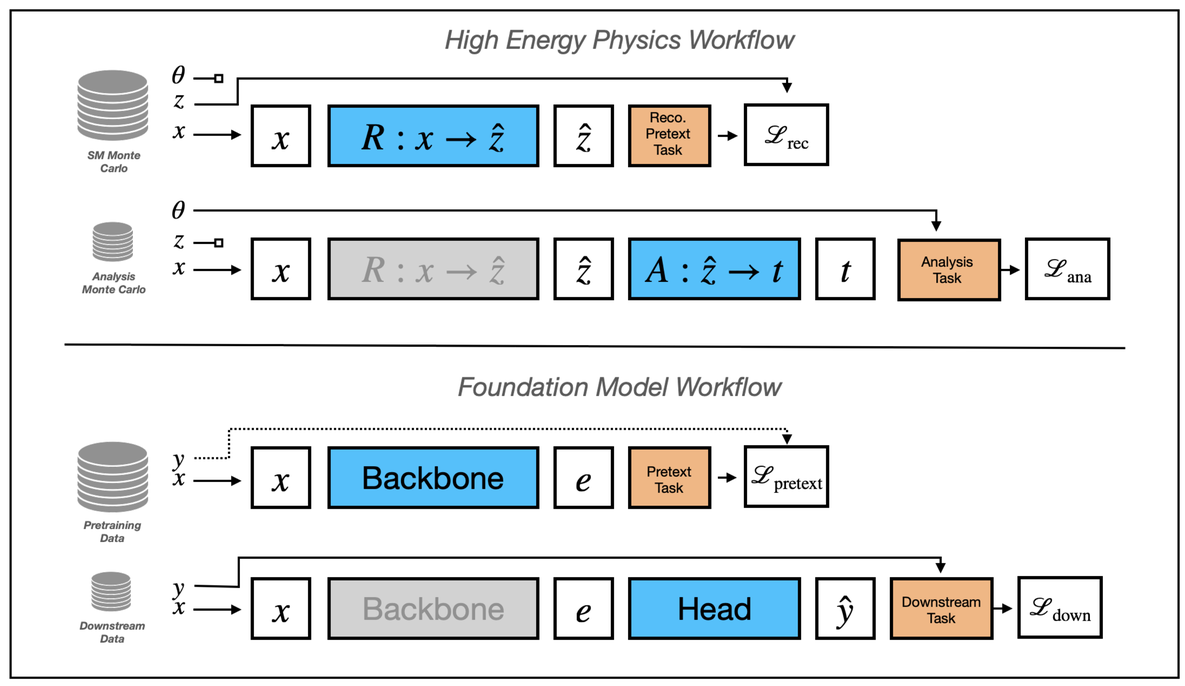
1. Xbb tagger infers whether a Large-R jet originated from a $H \to b\bar{b}$ decay from jet constituents. This + 4-momentum (HLF) provides a ***frozen*** Jet representation

2. Jets within the event are analyzed to perform a full event classification (S/B)

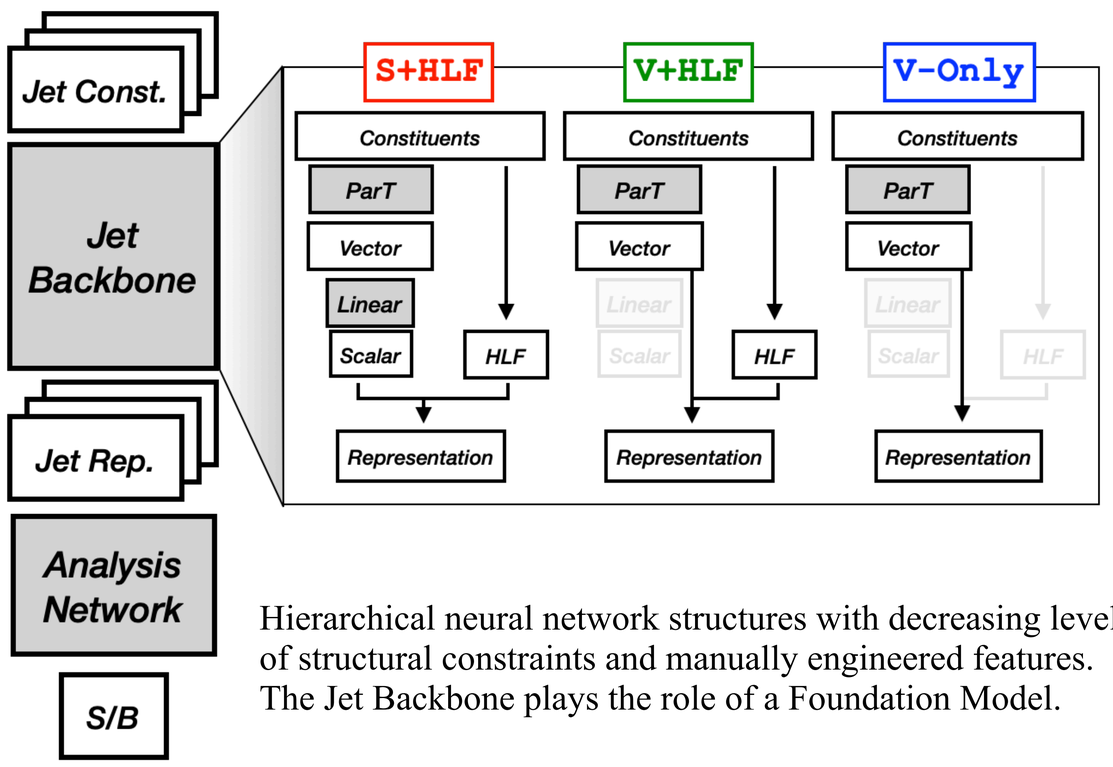


HEP in the Language of Foundation Models (FM)

Reconstruction plays the role of a backbone or foundation model yielding a general purpose representation of high-dimensional low-level data. The physics data analysis itself is a "head" that produces task-specific summary statistics.



Architecture

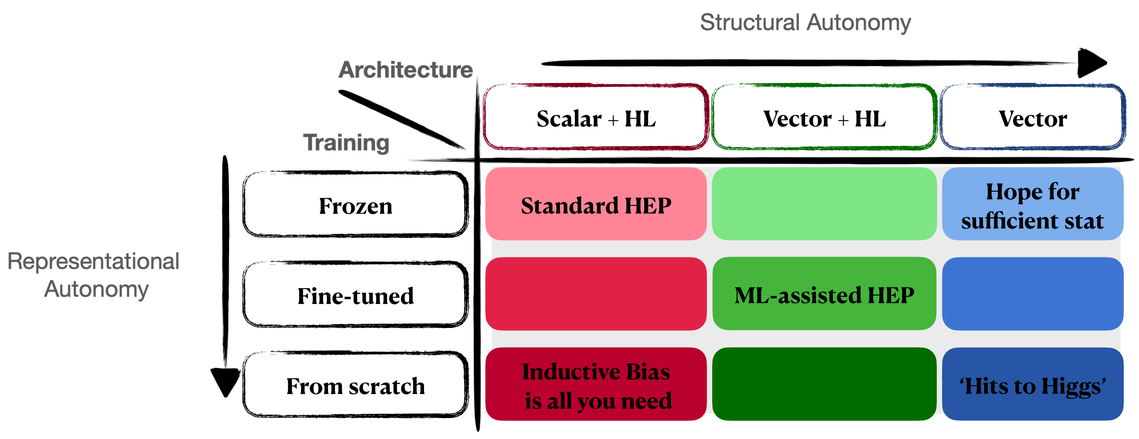


Hierarchical neural network structures with decreasing levels of structural constraints and manually engineered features. The Jet Backbone plays the role of a Foundation Model.

Do high-dim embeddings hold more (useful) info than Xbb+HL features?

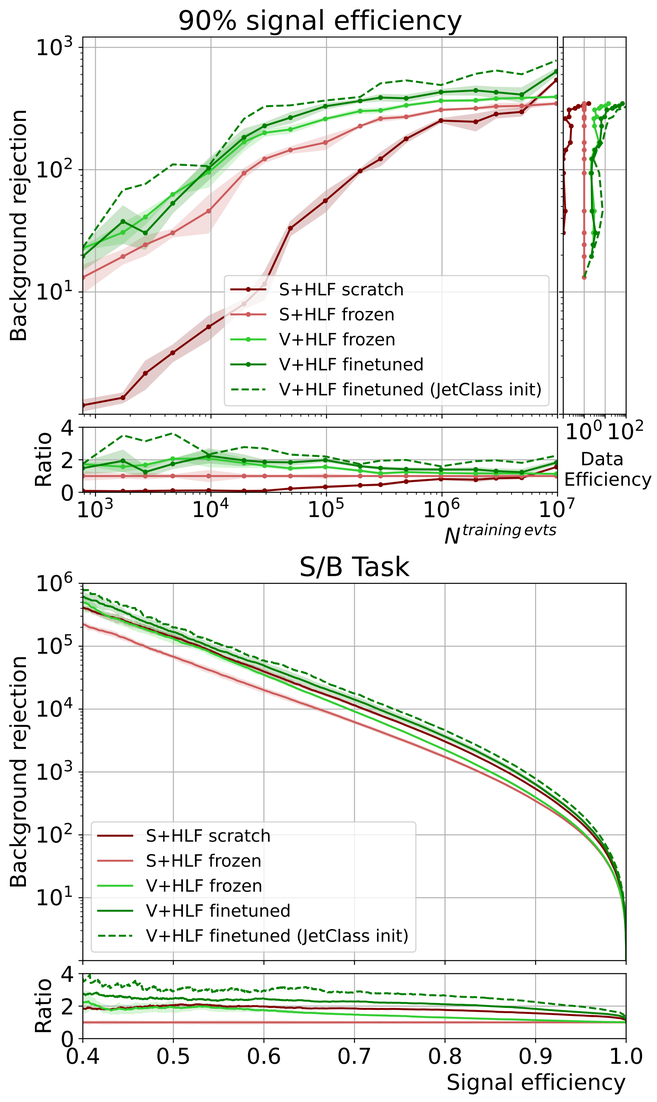
Training strategies

- Can we improve from the standard HEP (**frozen** and individually optimised tasks) paradigm?
- Is it useful to **pre-train** the Jet Backbone (our FM) on the Xbb pretext task and later **fine-tune** it on the downstream analysis task?
- Can we just train the whole model from **scratch**?



Results

Strategies from modern ML such as large-scale pretraining, finetuning, domain adaptation and high-dimensional embeddings (green curves) can lead to significant performance gains over the traditional HEP approach.



Data efficiency

Large (10x-100x) efficiency gain w.r.t. standard HEP, coming from high-dimensional embeddings and finetuning.

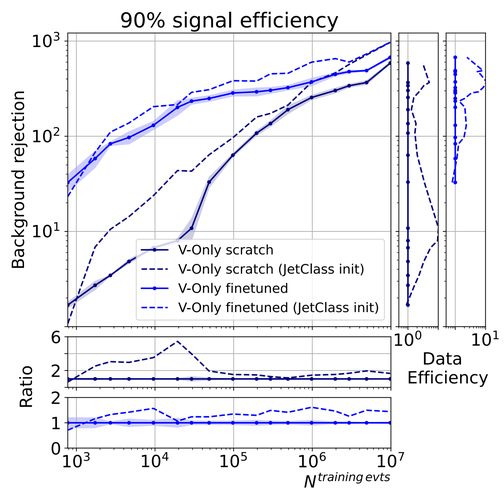
Training from scratch eventually surpasses the baseline model after enough training samples.

Performance gains

Clear performance hierarchy between training strategies: depending on the finetuned models, the gain in rejection can be as much as a factor of two larger than the frozen backbone.

Domain adaptation

Pre-training the backbone on a different data set (JetClass - 100M Jets) improves performance on downstream task (dashed lines).



References

[1]: Huilin Qu, Congqiao Li, and Sitian Qian, "Particle Transformer for Jet Tagging," (2022), arXiv:2202.03772

[2]: Duarte Javier, "Sample with jet, track and secondary vertex properties for hbb tagging ml studies. cern open data portal." (2019), DOI:10.7483/OPENDATA.CMS.JGJX.MS7Q.

Acknowledgements

LH and NH are supported by the Excellence Cluster ORIGINS, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-2094-390783311.