A Comprehensive Comparison of GNN Architectures for Jet Tagging

Daniel Murnane[1], Savannah Thais[2], Jason Wong[3]

[1]Lawrence Berkeley National Lab, [2]Princeton University, [3]UC Berkeley

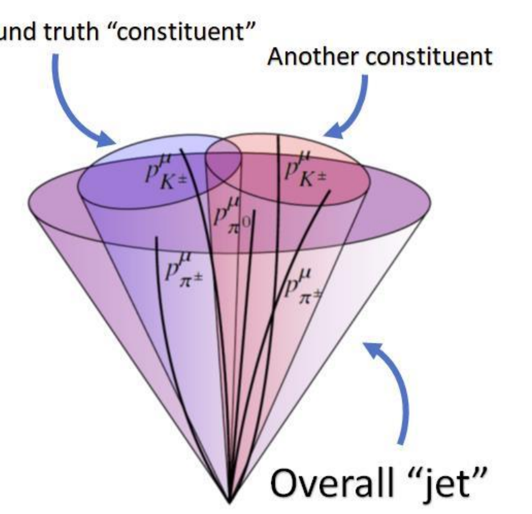
MOTIVATION

Jet tagging is a critical task in online and offline computing at the LHC

- Online trigger systems have strict latency requirements for full event processing: 4 μs for Level 1 and 200 ms for the High Level Trigger
- Reducing model size can decrease inference time, enabling ML-based tagging models to be used in experiment triggers
- Enforcing expected equivariance is a proposed technique to decrease model size while maintaining performance

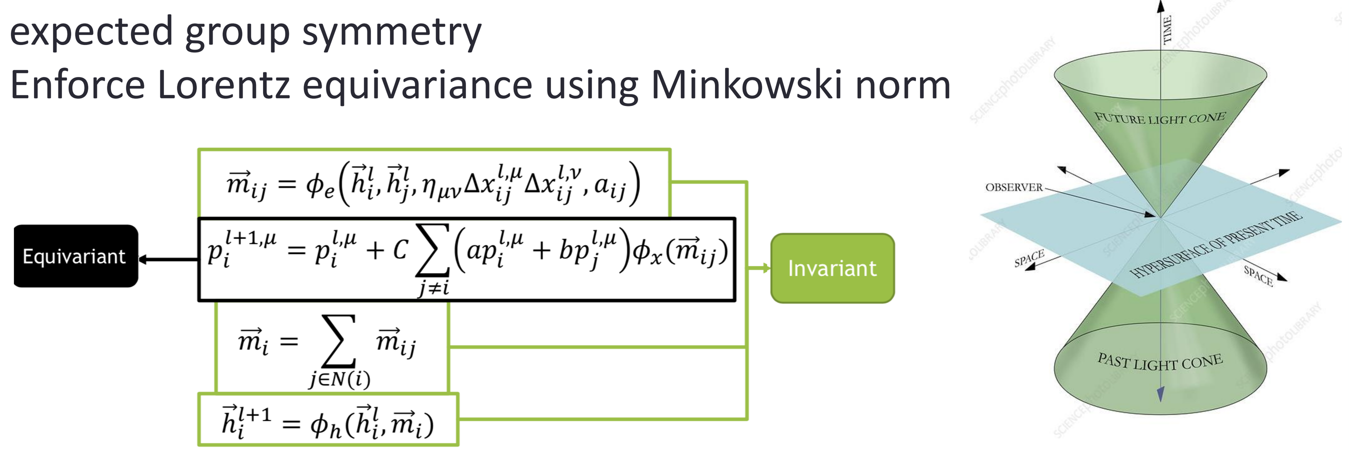
JET TAGGING

- Quarks and gluons produced in pp collisions at the LHC produce collimated sprays of particles called jets
- After jets are reconstructed from detector data, we need to identify what particle they originated from
- We focus on the problem of distinguishing top jets from light quark/gluon jets
- Use open-source 14TeV ATLAS dataset[1]
  - 1.2m training jets, 400k testing & validation jets
  - Selected by $p_T \in (550,650)$GeV, cluster $\Delta R < 0.8$, pseudorapidity $|\eta| < 2$
  - Max 200 constituents



EQUIVARIANT GNNS

- To enforce equivariance in a GNN the passed messages must be constructed of equivariant information
- Following EGNN[2] formulation, message equivariance is enforced by taking linear combinations of vectors and transforming them under the expected group symmetry
- Enforce Lorentz equivariance using Minkowski norm



LANDSCAPE OF TAGGERS

Several ML-based jet tagging architectures are already well studied[3]

Model	Accuracy	AUC	ϵ_B^{-1}	N_{params}	*ant*
ResNeXt	0.936	0.984	1122 ± 47	1.46M	0.0007
ParticleNet	**0.938**	**0.985**	$\mathbf{1298 \pm 46}$	498k	0.0026
EFP	0.932	0.980	384	1k	**0.384**
LGN	0.929	0.964	435 ± 95	4.5k	0.097

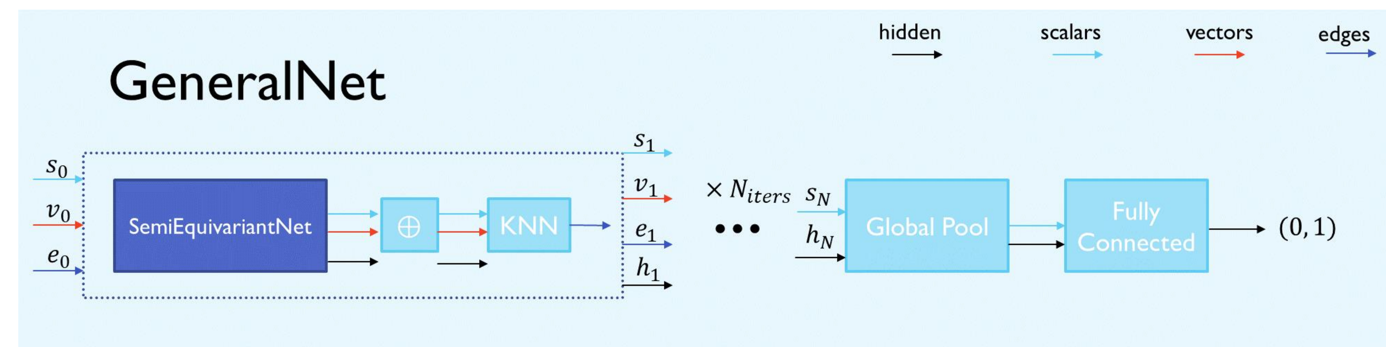
To study tradeoff between model accuracy and size, define the **ant factor**:

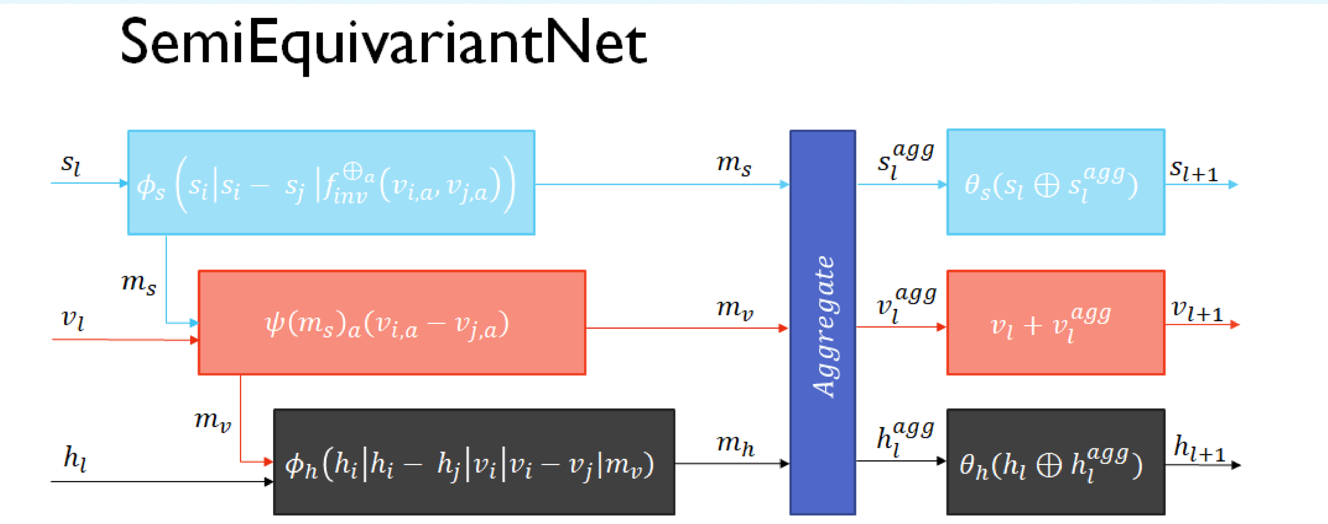
$$ant = \frac{accuracy}{model\ size} = \frac{\epsilon_B^{-1}}{N_{params}}$$

For background rejection rate $\frac{1}{\epsilon_B} = \frac{1}{fpr}$ at a particular signal efficiency

GNN ARCHITECTURE

In order to study the impact of different GNN components and behaviors on jet tagging we developed *GeneralNet,* which contains a convolution *SemiEquivariantNet.* This Interaction Network-like convolution allows scalar and vector objects to pass through while maintaining their group transformation properties. Hidden channels do not transform as a group representation.





- Scalar message: $m_s = \phi_s\left(s_i|s_i - s_j\,|f_{inv}^{\oplus_a}(v_{i,a}, v_{j,a})\right),$
  where $f_{inv}(v_{i,a}, v_{j,a}) = |v_{i,a} - v_{j,a}|^2 = -(v_{i,a}^0 - v_{j,a}^0)^2 + (v_{i,a}^1 - v_{j,a}^1)^2 + (v_{i,a}^2 - v_{j,a}^2)^2 + (v_{i,a}^3 - v_{j,a}^3)^2$
- Vector message: $m_v = \psi(m_s)_a(v_{i,a} - v_{j,a}),$ where $a = (1, \ldots, N_{heads})$
- Hidden message: $m_h = \phi_h(h_i, h_i - h_j, v_{i,a}, v_{i,a} - v_{j,a}, m_v)$
- Aggregation: $s^{agg} = \frac{1}{N}\Sigma_j^N(m_s)_{ij},$ $v^{agg} = \frac{1}{N}\Sigma_j^N(m_v)_{ij},$ $h^{agg} = \frac{1}{N}\Sigma_j^N(m_h)_{ij}$
- Node update: $s_{l+1} = \theta_s(s_l|s_l^{agg}),$ $v_{l+1} = v_l + v^{agg},$ $h_{l+1} = \theta_s(h_l|h_l^{agg})$

EXPERIMENTS

We systematically vary the different hyperparameters (HP) using Weights & Biases HP tuning API, over 1000 HP combinations

- Using Bayesian optimization with AUC score as figure of merit
- Trained on NERSC HPC systems (GPUS) for approx. 36 compute-days

Model Hyperparameters

Hyperparameter	Values
N edge layers $\phi s, \phi h$	∈[1,3]
N node layers θs, θh	∈[0,3]
N graph iterations	∈[1,6]
N scalar dimensions	∈[1,256]
N hidden dimensions	∈[1,256]
N attention heads ψ	∈[1, 256]
Shortcut	None, Skip, Concat
Activations	ReLU, SiLU, Tanh
Batch norm, Layer norm	True, False

Training Hyperparameters

Hyperparameter	Values
Graph construction	Static KNN, Dynamic KNN, Fully Connected
K neighbours	∈[3, 32]

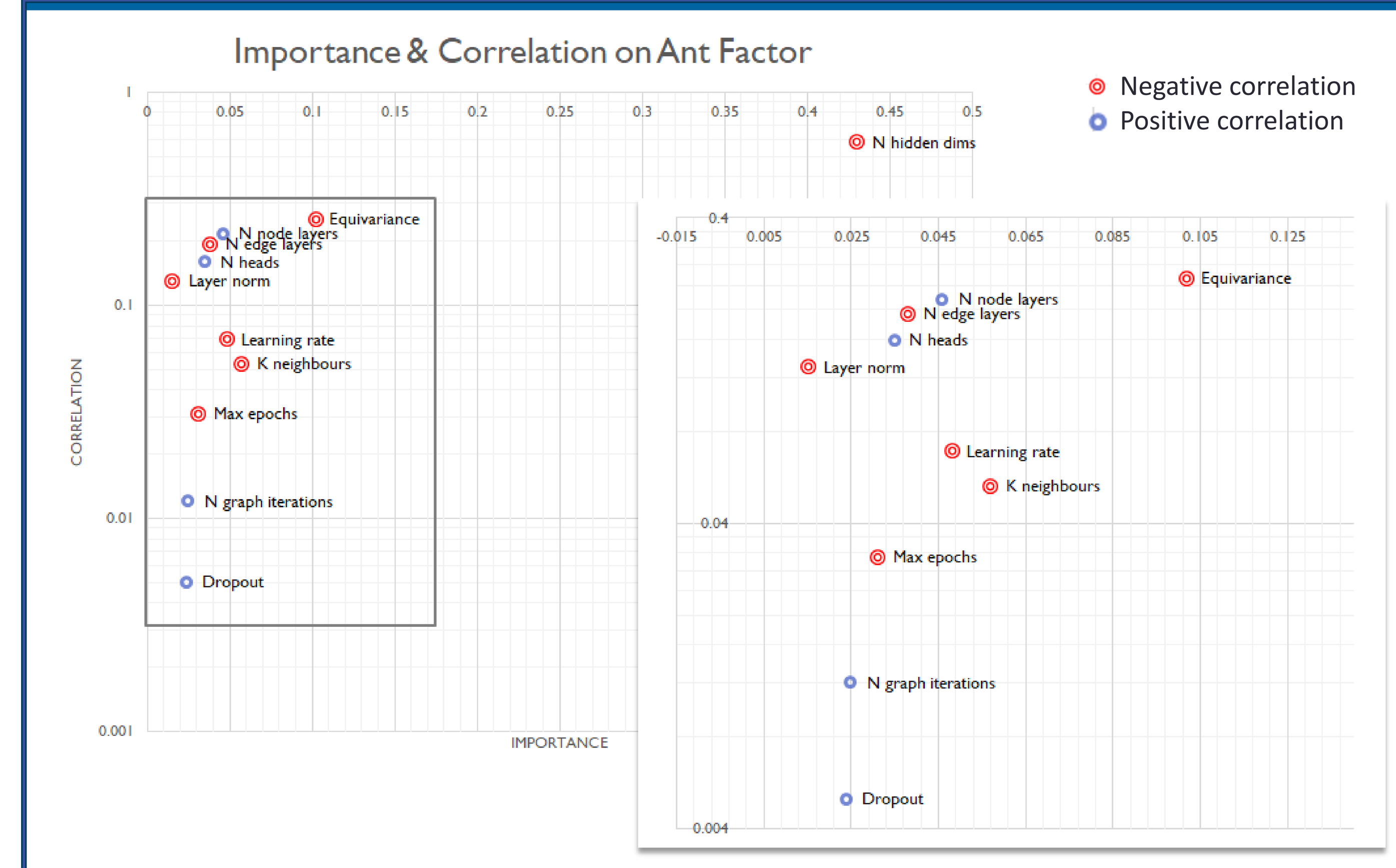
Graph Hyperparameters

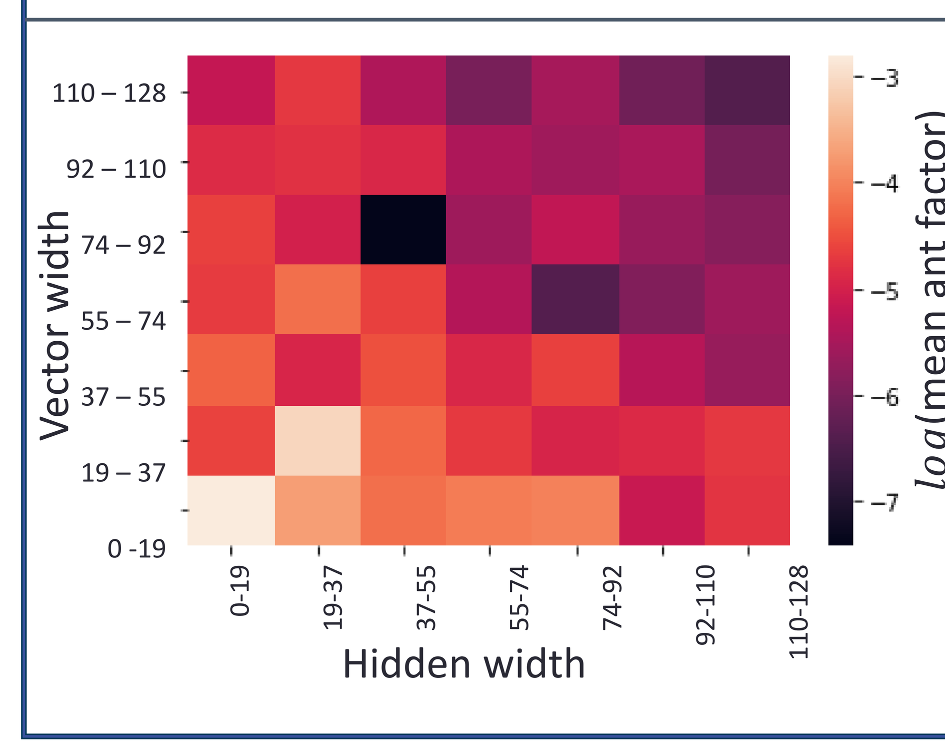
Hyperparameter	Values
Dropout	∈[0.01, 0.4]
Learning rate	∈[10−5, 10−2]
Max epochs	∈[10, 50]

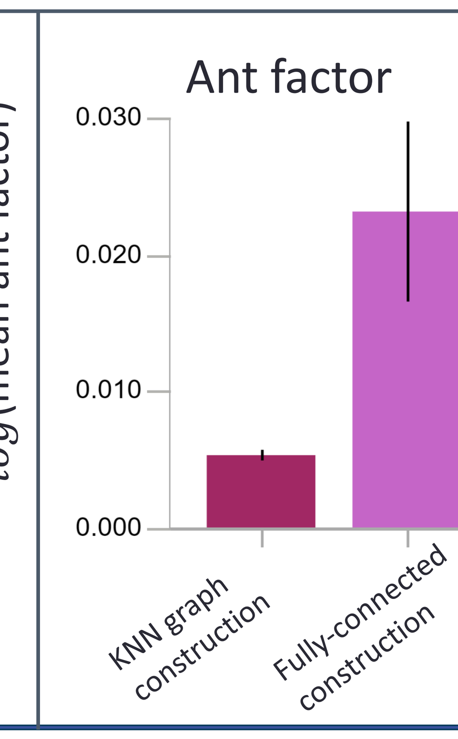
To understand which hyperparameters that the AUC and ant factor are most sensitive to we use built in W&B importance and correlation tools

- Correlation: linear correlation of AUC/ant factor vs hyperparameter
- Importance: library trains a random forest (RF) with the set of run HPs, with prediction goal of AUC or ant factor. RF naturally produces "variable importance" values

RESULTS







CONCLUSIONS

The GeneralNet with SemiEquivariant convolution attains SotA performance (AUC=0.9840), with AUC strongly correlated to network depth and number of hidden features. Importantly, by RF HP analysis, we see very little importance in number of message-passing iterations or KNN neighborhood size. However, very good performance (AUC=0.9834) is obtained with 20x smaller networks, by exploiting a narrow vector channel and a narrow hidden channel, and a fully connected graph. In general, best ant factors are obtained by combining narrow (<40) vector channels and narrow (<40) hidden channels. This suggests the network can quickly learn Lorentz-invariant physics with vector channels, while gaining last-mile expressiveness from hidden channels. It remains to study the physical phenomena learned by this combination of semi-equivariance, and follow-up studies are currently using interpretability tools for this.

REFERENCES

1. https://zenodo.org/record/2603256#.YXoI3Z5KjEZ
2. VG Satorras, E Hoogeboom, and M Welling, "E(N) Equivariant Graph Neural Networks" (2021)
3. A Bogatskiy et al, "Lorentz Equivariant Neural Networks for Particle Physics" (2020)
4. H Qu and L Gouskos, "ParticleNet: Jet Tagging via Particle Clouds" (2020)

Contact:
Daniel Murnane,
dtmurnane@lbl.gov

This work is supported by IRIS-HEP through the U.S. National Science Foundation (NSF) under Cooperative Agreement OAC-1836650 and by Department of Energy grant DE-SC0007968.