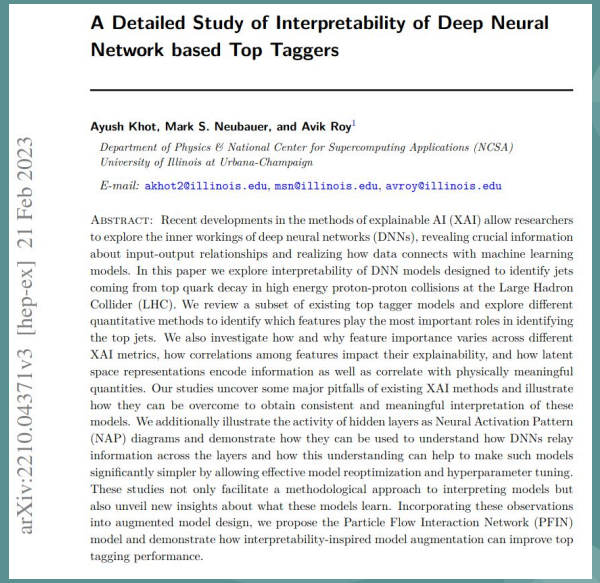
Interpretability Inspires: Explainable AI for DNN Top Taggers

Avik Roy, Ayush Khot, Mark Neubauer
University of Illinois at Urbana-Champaign



arXiv:2210.04371v3 [hep-ex] 21 Feb 2023

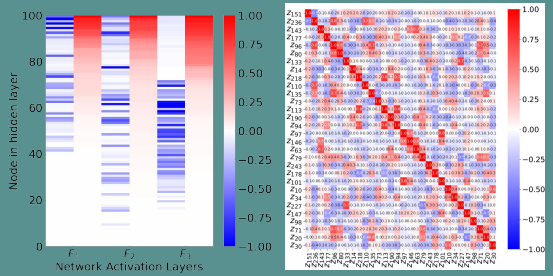
A Detailed Study of Interpretability of Deep Neural Network based Top Taggers

Ayush Khot, Mark S. Neubauer, and Avik Roy[1]

Department of Physics & National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign

E-mail: akhot2@illinois.edu, msn@illinois.edu, avroy@illinois.edu

Abstract: Recent developments in the methods of explainable AI (XAI) allow researchers to explore the inner workings of deep neural networks (DNNs), revealing crucial information about input-output relationships and realizing how data connects with machine learning models. In this paper we explore interpretability of DNN models designed to identify jets coming from top quark decay in high energy proton-proton collisions at the Large Hadron Collider (LHC). We review a subset of existing top tagger models and explore different quantitative methods to identify which features play the most important roles in identifying the top jets. We also investigate how and why feature importance varies across different XAI metrics, how correlations among features impact their explainability, and how latent space representations encode information as well as correlate with physically meaningful quantities. Our studies uncover some major pitfalls of existing XAI methods and illustrate how they can be overcome to obtain consistent and meaningful interpretation of these models. We additionally illustrate the activity of hidden layers as Neural Activation Pattern (NAP) diagrams and demonstrate how they can be used to understand how DNNs relay information across the layers and how this understanding can help to make such models significantly simpler by allowing effective model reoptimization and hyperparameter tuning. These studies not only facilitate a methodological approach to interpreting models but also unveil new insights about what these models learn. Incorporating these observations into augmented model design, we propose the Particle Flow Interaction Network (PFIN) model and demonstrate how interpretability-inspired model augmentation can improve top tagging performance.

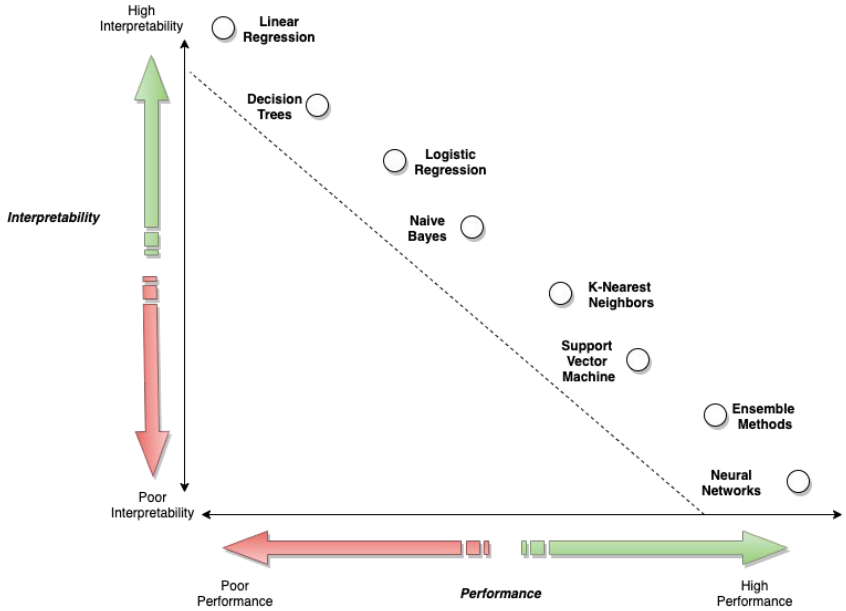


Results from arxiv: 2210.04371
Git repo: https://github.com/FAIR4HEP/xAI4toptagger/

CHEP, Norfolk, VA
May 09, 2023

Explainable AI (XAI) for Top Tagging

- **Top tagging**: Identify jets originating from top quarks amid background (e.g. QCD)
- We want to answer a few fundamental questions about model explanations-
  - What features are important?
  - Are interpretations consistent across methods? If not, why?
  - How information travels within a model?
  - What do networks learn in their latent spaces?
- Studies done with benchmark [top-tagging dataset] (includes 1M top and 1M QCD jets)



source: AWS docs

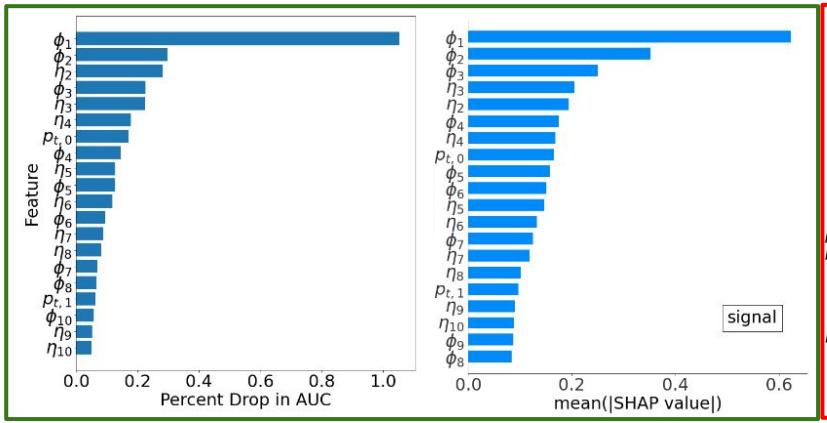
Methods of Explainability

- Occlusion test with ΔAUC score
  - Find feature ranking based on replacing certain features with their mean values and calculating the change in model's ROC-AUC score
- SHAP scores [link]:
  - Use the model-agnostic Kernel SHAP approach to identify the weighted marginal contribution of each feature
- Layer-wise Relevance Propagation (LRP) [link]:
  - Back propagates the score from the final output layer to original inputs using a linear redistribution
- Neural Activation Pattern (NAP diagram):
  - Relative Neural Activity (RNA) at each node and visualises information pathways along with model's sparsity

	ΔAUC	SHAP	LRP	RNA/NAP
Scalability in input dimension	✗	✗	✓	✓
Local explanation	✗	✓	✓	✗
Global explanation	✓	✓	✓	✓
Requires Forward Propagation	✓	✓	✓	✓
Requires Backward Propagation	✗	✗	✓	✗
Susceptible to spurious correlations	✓	✓	✓	✗
Addresses Model Complexity	✗	✗	✗	✓
Requires Retraining	✗	✗	✗	✗

Feature Importance in TopoDNN

- Simplest DNN architecture, implemented with an MLP with multiple hidden layers
- Uses preprocessed p_t, η, φ of top 30 (p_t ordered) jet constituents- zero padding for missing entries



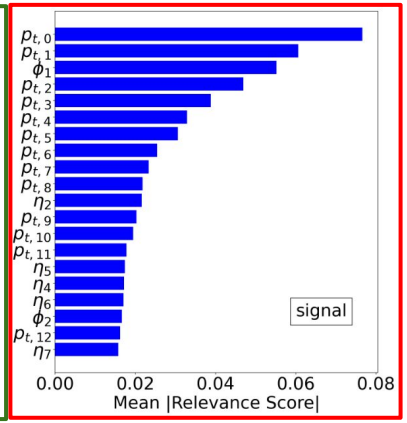
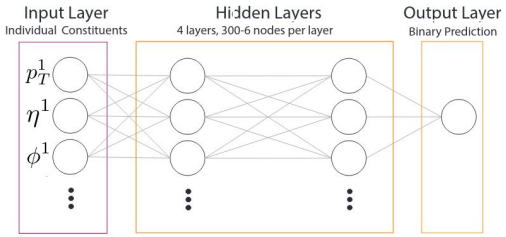
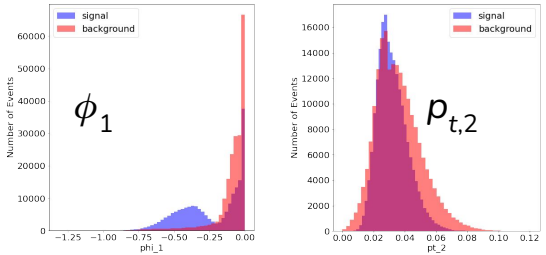


Image from 1704.02124



Baseline Architecture and Performance	
N_{in}, N_{out}	90, 1
Hidden Layers	(300, 102, 12, 6)
Accuracy, AUC	91.6%, 0.971



Why are results from LRP so different and assign large scores to non-expressive features?

Making relevances relevant: Differential Relevance

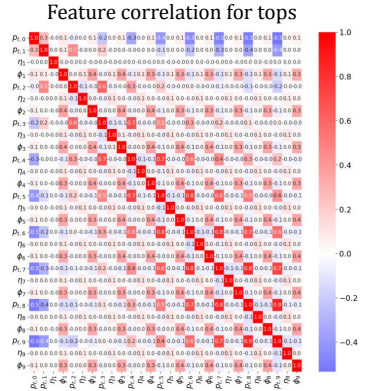
$$f(\vec{x}) = \sum_i r(x_i) \approx f(\vec{x}_{\setminus k}) + \frac{\partial f}{\partial x_k}(x_k - \bar{x}_k)$$

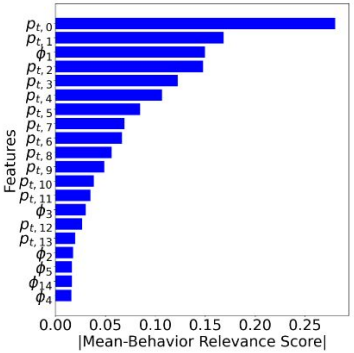
$$f(\vec{x}_{\setminus k}) = \sum_{i \neq k} r(x_i) + r(\bar{x}_k)$$

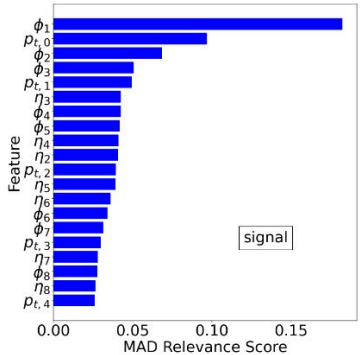
Differential relevance
(first order approximation)

Mean-behavior relevance

- **When features are uncorrelated (or weakly correlated)**, calculate mean-behavior relevance by simply replacing all features by their mean value and then calculating their relevances
- Differential relevance is more exact, determined by simply calculating the deviation in model's output when a particular feature is replaced by its mean value







MAD relevance:
Mean Absolute Differential Relevance

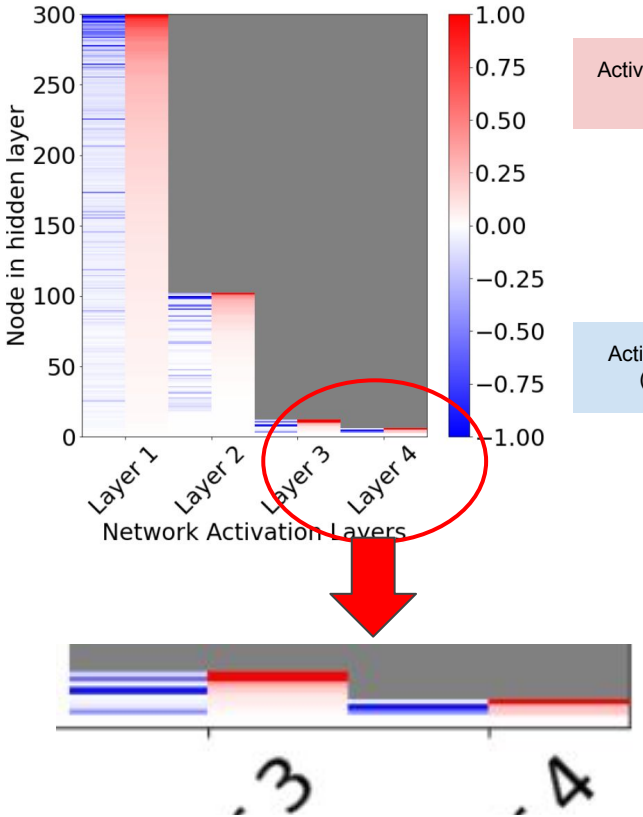
Has a stronger resemblance with the SHAP scores since this takes the "deviations" into account. (Actually, diff. Rel. is one of the leading terms that contribute to SHAP score)

Neuron Activation Patterns (NAPs)

- Understanding the model's inner workings- detect internal disentanglements, context-aware neural pathways, hyperparameter reoptimization
- Define Relative Neural Activity (RNA) score for different nodes within a layer

$$\mathrm{RNA}(j,k;\mathcal{S}) = \frac{\sum_{i=1}^{N} a_{j,k}(s_i)}{\max_j \sum_{i=1}^{N} a_{j,k}(s_i)}$$

- Observations:
  - The model is very sparse
  - The information pathways for jet classes are disentangled by layer 3, layer 4 is kind of redundant
- Retrained the model with (120,40,6) hidden nodes, got the same performance



Activation for signal (top jets)

Activation for bkg (QCD jets)

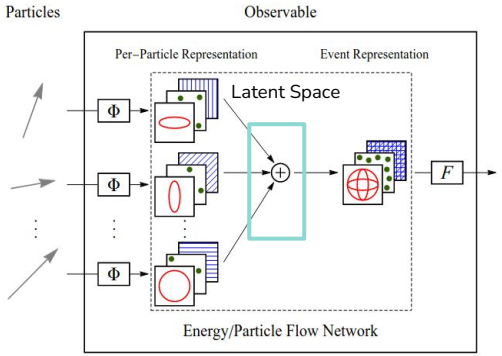
Latent Space in Particle Flow Network (PFN)

- Deep-set architecture, invariant under permutation of constituents

$$\mathrm{PFN} = F\left(\sum_{i=0}^{N-1} \Phi\left(p_i\right)\right)$$

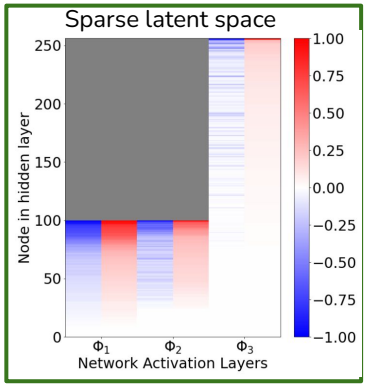
- Use MLPs to approximate the non-linear functions Φ and *F*
- Obtain **latent space representation** for jet level observables

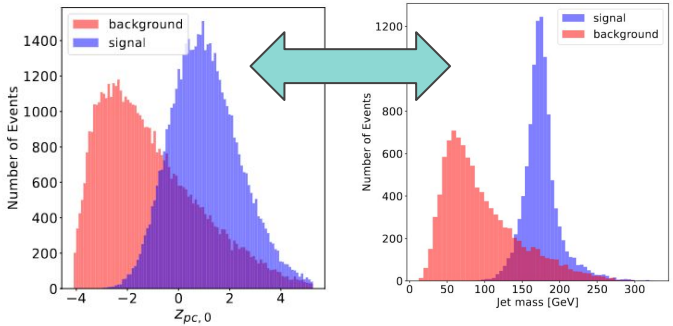
Image from [1810.05165]



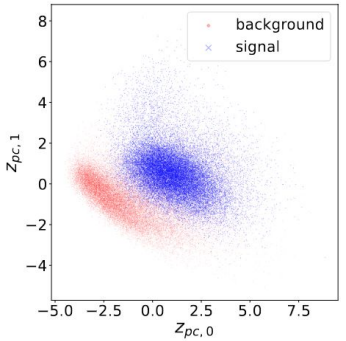
Baseline Architecture and Performance	
N_{in}, N_{out}	Φ: 3,256 *F*: 256,2
Layers	Φ: (3,100,100,256) F: (256,100,100,100,2)
Accuracy, AUC	92.8%, 0.980

Latent space learns to mimic physical observables



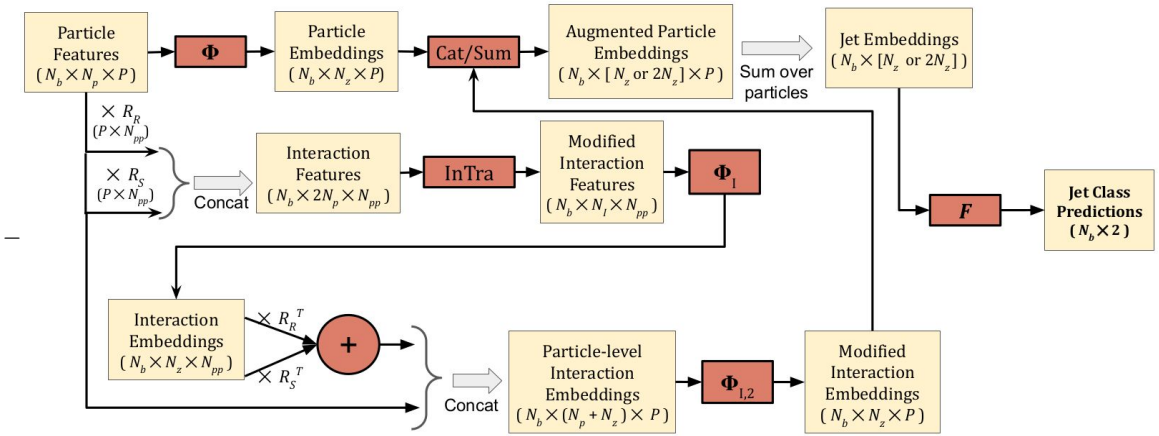


Jet class information is encoded in the correlation structure of the latent spaces



Interpretability Inspired Model
The Particle Flow Interaction Network

- What did we learn from the XAI studies of TopoDNN and PFN?
  - PFN is limited by not considering inter-particle interactions is considered - *room for improvement!*
  - Latent space for PFN is sparse – scope for model simplification
- Augment the PFN model with a Graph-net called Interaction Network (IN)
- Models the pairwise particle interaction in the latent space



Interpretability Inspired Model
The Particle Flow Interaction Network (PFIN)

Model Hyperparameters	
Number of constituents, N_p	60
Nodes in Φ Network	(100,100,64)
Nodes in Φ_I Network	(128,128,64)
Nodes in $\Phi_{I,2}$ Network	(128,128,64)
Nodes in F Network	(64,100,100)
Latent space dimension	64 (s), 128 (c)
Number of Parameters	97k (s), 101k (c)
Performance Metrics	
ROC-AUC	0.9839 (s), 0.9838 (c)
Accuracy	0.937 (s), 0.937 (c)
Background Rejection Rate ($1/\epsilon_B$)	1041 (s), 1030 (c)

Number of constituents reduced to 60 (from 200) since only the most energetic constituents show up as important features

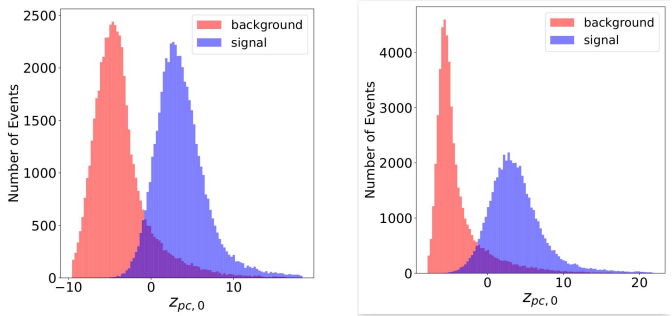
Network architectures are inspired by the NAP diagrams for the PFN model

Outperforms both PFN and the IN models, comparable with ResNext and ParticleNet with a much smaller number of parameters and faster convergence

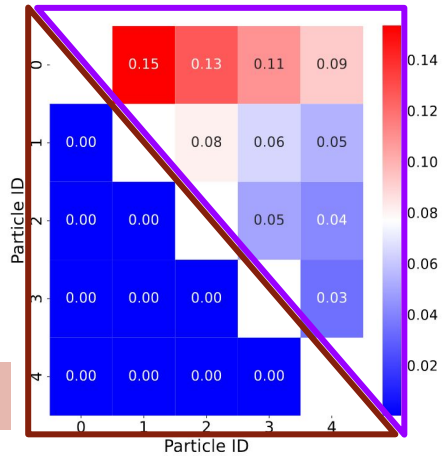
Interpreting PFIN: the Latent Space and the Interaction Features

- PFIN latent space shows a much stronger correlation with the jet mass and the subjettiness variables
- We can investigate the importance of pairwise particle interactions using MAD relevance of probability scores
- Inter-particle interactions play a significant role in top jet identification compared to QCD jets

Latent space learns to mimic physical observables



Average impact on classification probability



top jets

QCD jets

Lessons Learned and Outlook

- Just like models themselves, *one size does not fit all* for model interpretation
- Model explanations can be tricky and unreliable when-
  - models have highly correlated inputs
  - models that concurrently treat categorical and continuous features
  - models whose inputs span over multiple orders
- RNA scores and NAP diagrams reveal important insight into model's desired complexity, can we use them for *in-situ* model optimization?
- Latent spaces are interesting- can they mimic physical features in more general settings (e.g. in multi-class classification) ?
- Interpreting more complex models like graph nets, transformers etc. may require even better techniques

Mark Neubauer

Ayush Khot