Interpretability of Top Taggers

Avik Roy
University of Illinois at Urbana-Champaign

Nov 15, 2022

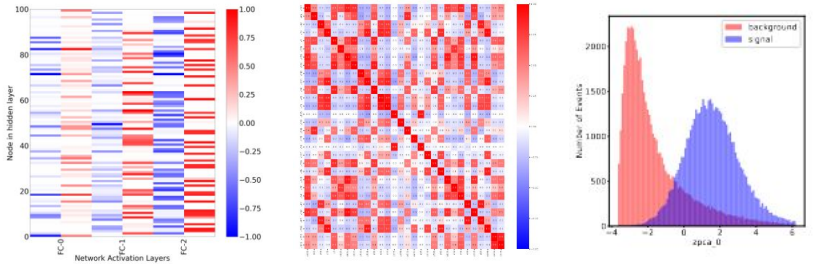
A Detailed Study of Interpretability of Deep Neural Network based Top Taggers

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) methods allow us 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 for identifying jets coming from top quark decay in the high energy proton-proton collisions at the Large Hadron Collider (LHC). We review a subset of existing such 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 feature correlations 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 us to make such models significantly simpler by allowing effective model reoptimization and hyperparameter tuning. While the primary focus of this work remains a detailed study of interpretability of DNN-based top tagger models, it also features state-of-the art performance obtained from modified implementation of existing networks.



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

Introduction to Top Tagging

- Identify (boosted) jets originating from top quarks amid background from QCD processes
- One of the major benchmark problems in ML landscape within HEP-ex
- Many models have been developed over the years
- A good review of these models can be found in this paper: https://scipost.org/SciPostPhys.7.1.014/pdf
  - A very good read for an overview of top tagging models, comparative performance
  - Made available the benchmark top-tagging dataset

SciPost Phys. 7, 014 (2019)

The Machine Learning landscape of top taggers

Gregor Kasieczka[1⋆], Tilman Plehn[2†], Anja Butter[2], Kyle Cranmer[3], Dipsikha Debnath[4], Barry M. Dillon[5], Malcolm Fairbairn[6], Darius A. Faroughy[5], Wojtek Fedorko[7], Christophe Gay[7], Loukas Gouskos[8], Jernej F. Kamenik[5,9], Patrick T. Komiske[10], Simon Leiss[1], Alison Lister[7], Sebastian Macaluso[3,4], Eric M. Metodiev[10], Liam Moore[11], Ben Nachman[12,13], Karl Nordström[14,15], Jannicke Pearkes[7], Huilin Qu[8], Yannik Rath[16], Marcel Rieger[16], David Shih[4] Jennifer M. Thompson[2], and Sreedevi Varma[6]

1 Institut für Experimentalphysik, Universität Hamburg, Germany
2 Institut für Theoretische Physik, Universität Heidelberg, Germany
3 Center for Cosmology and Particle Physics and Center for Data Science, NYU, USA
4 NHECT, Dept. of Physics and Astronomy, Rutgers, The State University of NJ, USA
5 Jozef Stefan Institute, Ljubljana, Slovenia
6 Theoretical Particle Physics and Cosmology, King's College London, United Kingdom
7 Department of Physics and Astronomy, The University of British Columbia, Canada
8 Department of Physics, University of California, Santa Barbara, USA
9 Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
10 Center for Theoretical Physics, MIT, Cambridge, USA
11 CP3, Universitéxx Catholique de Louvain, Louvain-la-Neuve, Belgium
12 Physics Division, Lawrence Berkeley National Laboratory, Berkeley, USA
13 Simons Inst. for the Theory of Computing, University of California, Berkeley, USA
14 National Institute for Subatomic Physics (NIKHEF), Amsterdam, Netherlands
15 LPTHE, CNRS & Sorbonne Université, Paris, France
16 III. Physics Institute A, RWTH Aachen University, Germany

⋆ gregor.kasieczka@uni-hamburg.de, † plehn@uni-heidelberg.de

Abstract

Based on the established task of identifying boosted, hadronically decaying top quarks, we compare a wide range of modern machine learning approaches. Unlike most established methods they rely on low-level input, for instance calorimeter output. While their network architectures are vastly different, their performance is comparatively similar. In general, we find that these new approaches are extremely powerful and great fun.

Our Target

- We want to answer a few fundamental questions about model interpretability-
  - Are interpretations consistent across methods? If not, why?
  - How information travels within a model?
  - What do networks learn in their latent spaces?
- Round 1: starting with the simpler models- explore interpretability without convoluting it with architectural complexity of unorthodox models
- Models we are exploring now:
  - TopoDNN: MLP trained on p_T, η, φ of constituent particles
  - Particle-Flow network: Trained on p_T, η, φ of constituents with a deepset architecture (permutation invariant)

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 fit an MSE-minimizing linear regression that assigns additive scores to each feature to explain the model's output
- 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):
  - Calculates Relative Neural Activity (RNA) at each node and visualises information pathways along with model's sparsity

TopoDNN

- Simplest DNN architecture, implemented with an MLP with multiple hidden layers
- Uses p_T, η, φ of top 30 (p_T ordered) jet constituents- zero padding for missing entries
- Data is pre-processed to
  - align the highest-p_T constituent along (0,0) in η-φ and
  - align the second highest- constituent along the negative φ axis
  - scale the p_T values by 1/1700

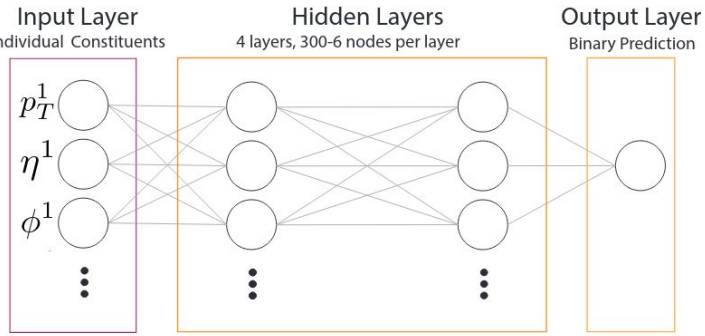
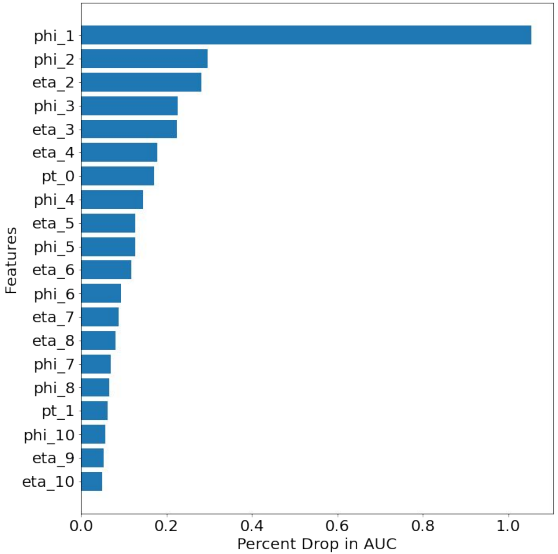


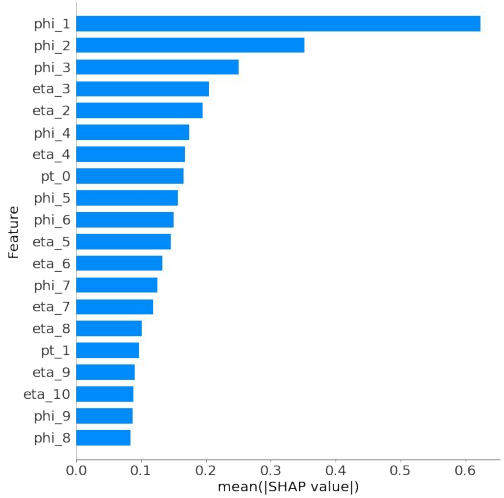
Image from 1704.02124

Baseline Architecture	
N_{in}	90 (= 3 x 30)
N_{out}	1
Hidden Layers	(300, 102, 12, 6)
Accuracy	91.6%
ROC-AUC	97.1%

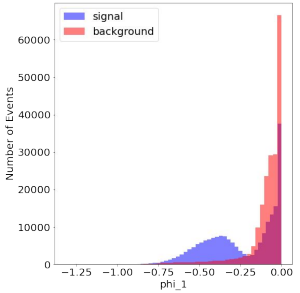
Results from TopoDNN

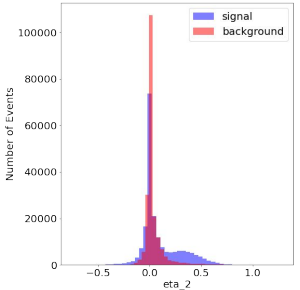


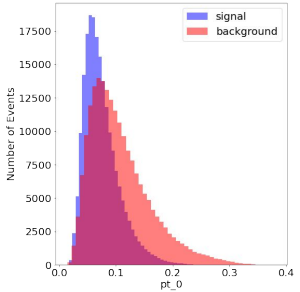
ΔAUC scores



SHAP scores
(signal/top jets)



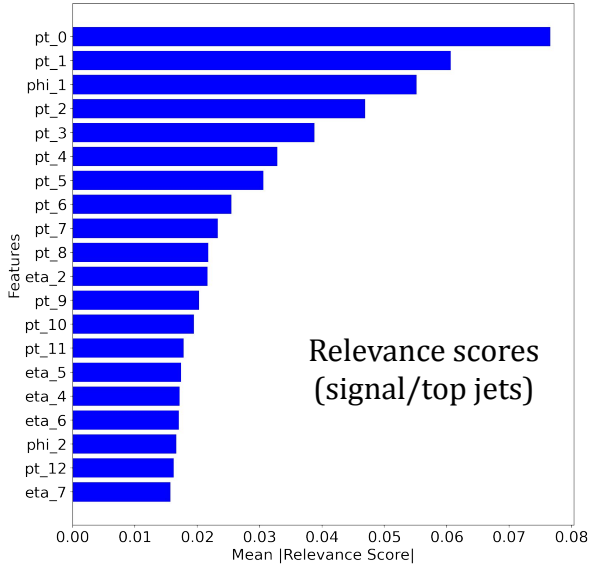


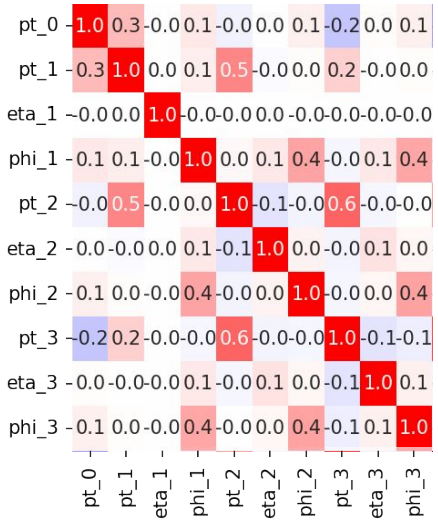


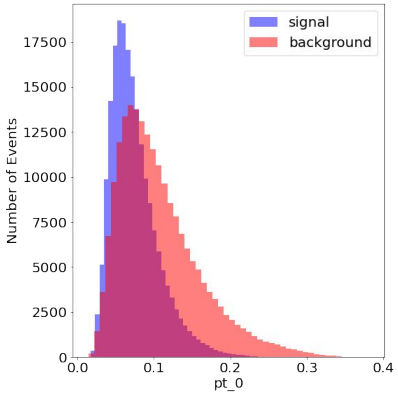
Results from TopoDNN

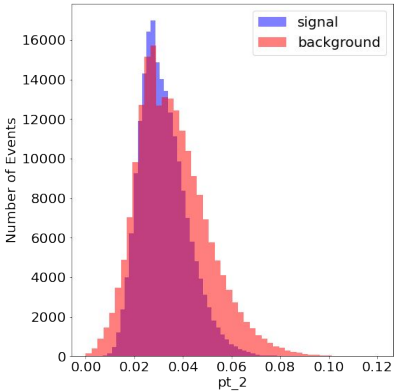
highest-ranked features are very different with LRP

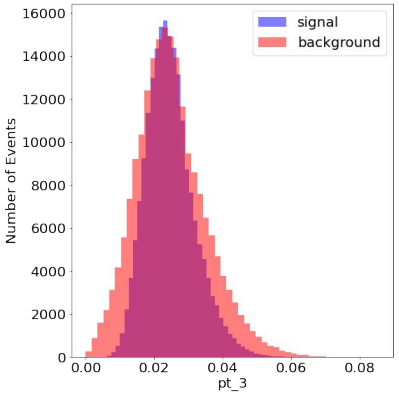
It is also counter-intuitive, pT's are minimally expressive variables, should not be considered as the most important features











Understanding the LRP results

- Why LRP assigns large relevance to *intuitively* less important features?
- Why SHAP and ΔAUC are similar and “very” different from LRP?

SHAP and ΔAUC calculate “Deviation” from mean-behavior

They represent the impact of including the true value of a feature compared to the mean or an *uninformative* value

LRP score includes mean-behavior relevance!!

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

$$\vec{x}_{\setminus k} = \vec{x} \setminus \{x_k\} \bigcup \{\mathbf{E}(X_k)\}$$

Disentangling relevance scores

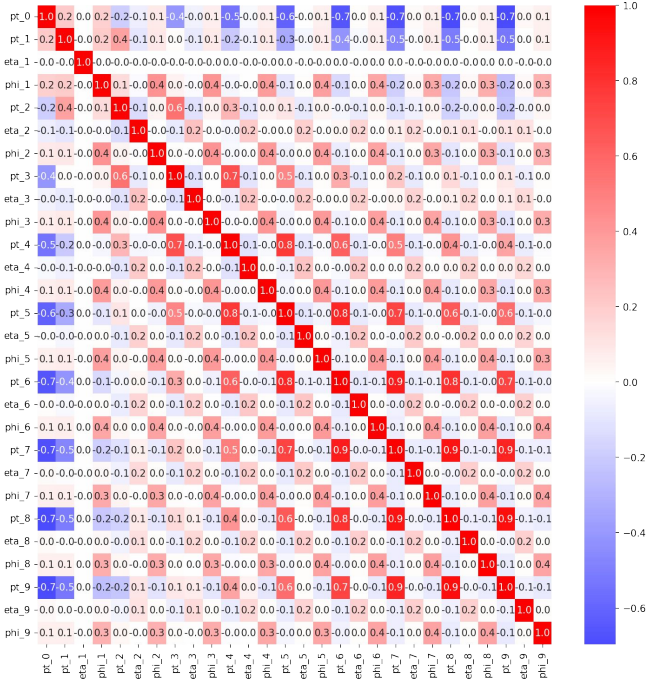
$$f(\vec{x}) = \sum_i r(x_i) \approx f(\vec{x}_{\setminus k}) + \frac{\partial f}{\partial x_k}(x_k - \bar{x}_k)$$

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

Differential relevance

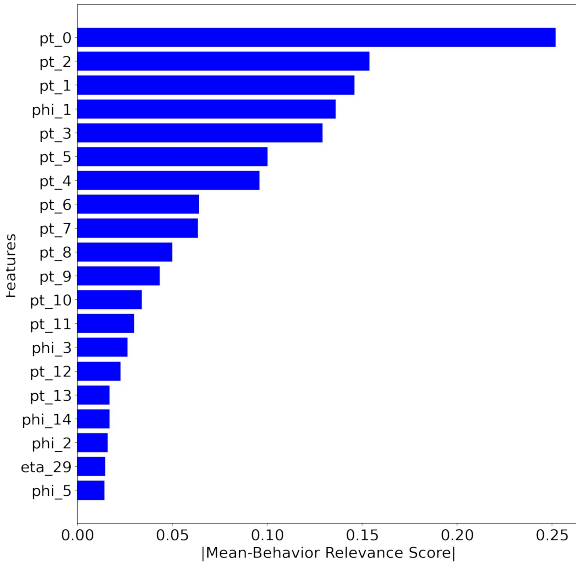
Mean-behavior relevance

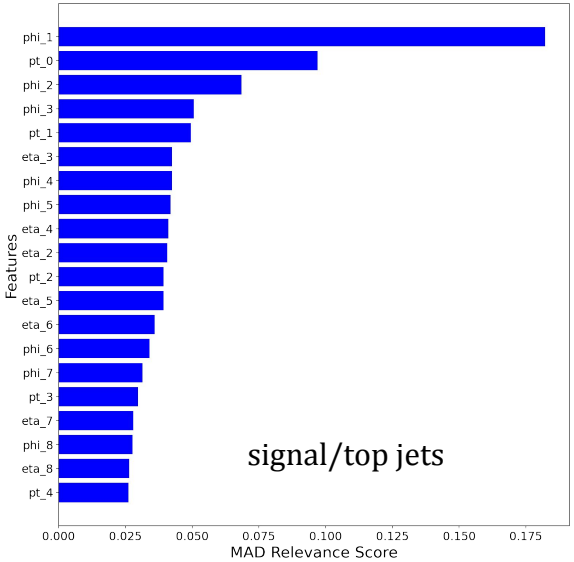
- **When features are uncorrelated (or weakly correlated)**, calculate mean-behavior relevance by simply replacing all features by their mean value and then calculating their relevances
- Differential relevance is more exact, determined by simply calculating the deviation in model's output when a particular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



Correlation among features for QCD jets

A better explainer: Differential Relevance





MAD Relevance = Mean Absolute Differential Relevance

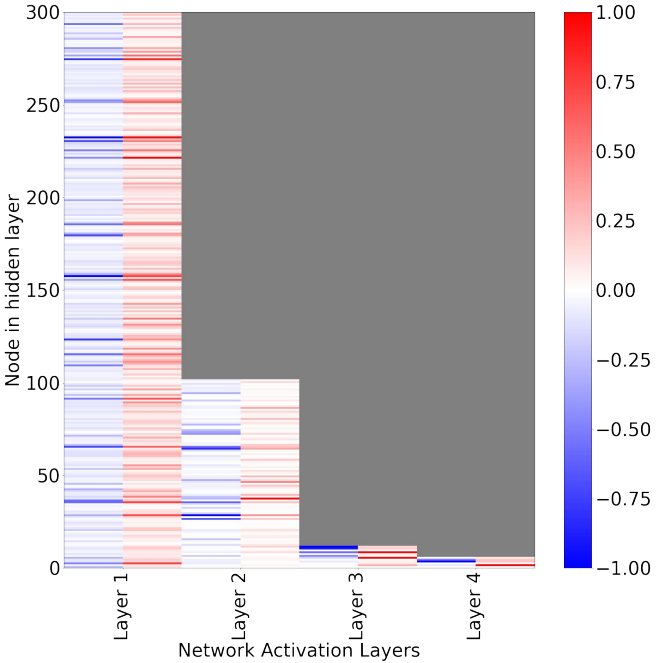
Diff. Rel. is a significantly better heuristic!

- Much easier to calculate, only need forward propagation
- An exact estimator
- Recovers physically intuitive explanations

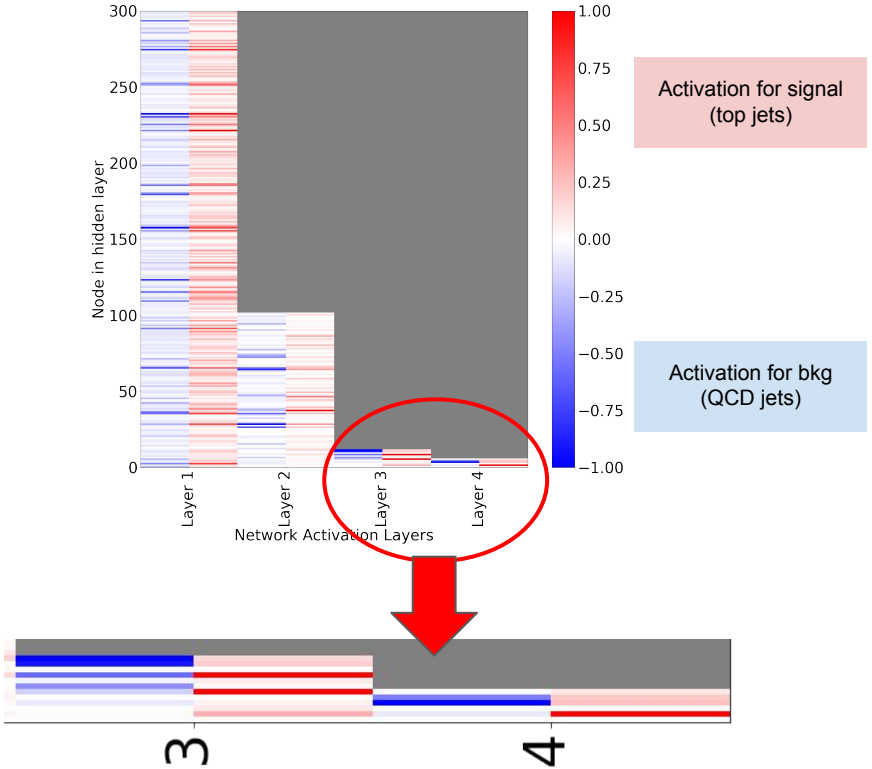
Neuron Activation Patterns (NAPs)

- Feature importance metrics don't reveal any information about the model's inner workings
- Understanding the model's inner workings help with hyperparameter reoptimization
- To see how the hidden layers respond to input data, we look at the Relative Neural Activity (RNA) score for different nodes within a layer

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



NAP Diagram for TopoDNN



- RNA scores of QCD jets mapped as negative numbers for simultaneous visualization
- Sparsity measured as the fractional number of nodes with RNA < 0.2
- The model is very sparse, Sparsity = 0.70
- 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- sparsity reduced to 0.59**

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
- Preprocess data to
  - scale constituent p_T by sum of constituent p_T
  - Subtract jet η, φ from constituents

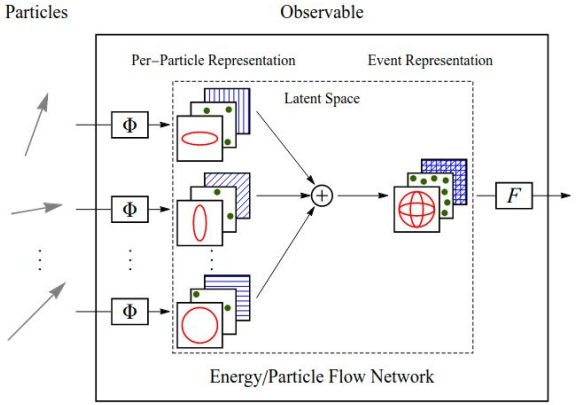


Image from 1810.05165

Baseline Architecture	
Ninput	Φ: 3, *F:* 256
Noutput	Φ: 256, *F:* 2
Layers	Φ: (3,100,100,256) F: (256,100,100,100,2)
Accuracy	97.7%
ROC-AUC	99.7%

A Note of PFN's Performace

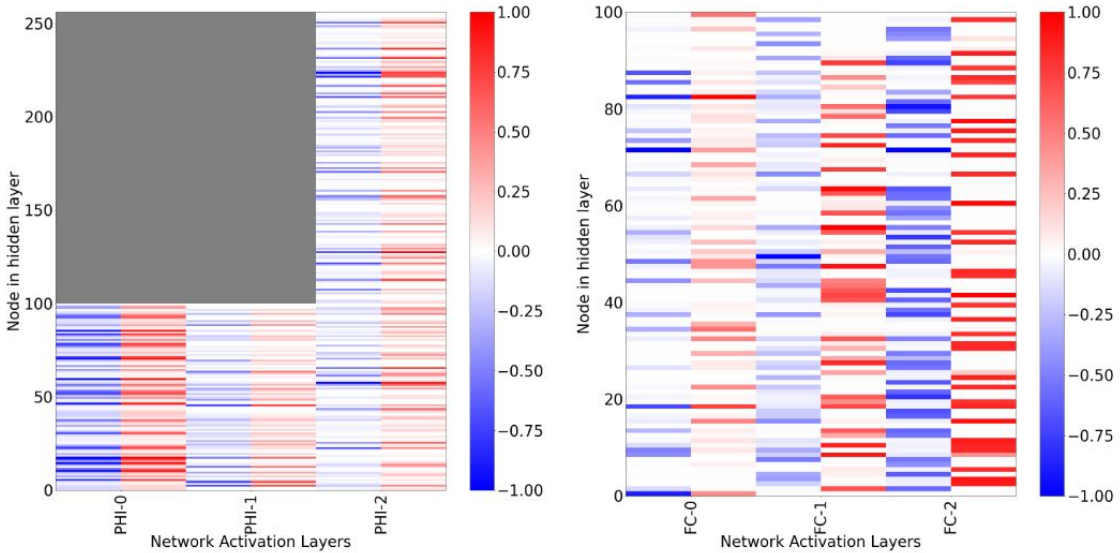
- Performance of the PFN model is reported in this seminal review article: *The Machine Learning Landscape of Top Taggers*
- We found their implementation to be significantly underperforming
- The difference probably comes from the choice of particle-level inputs: we chose to preprocess the inputs prescribed in the actual implementation proposed in 1810.05165

Quoted from SciPost Phys. 7, 014 (2019)	AUC	Acc
LBN [28]	0.981	0.931
LoLa [31]	0.980	0.929
LDA [63]	0.955	0.892
Energy Flow Polynomials [30]	0.980	0.932
Energy Flow Network [32]	0.979	0.927
Particle Flow Network [32]	0.982	0.932

Results from our implementation	
Accuracy	**97.7%**
ROC-AUC	**99.7%**

NAP Diagrams for TopoDNN

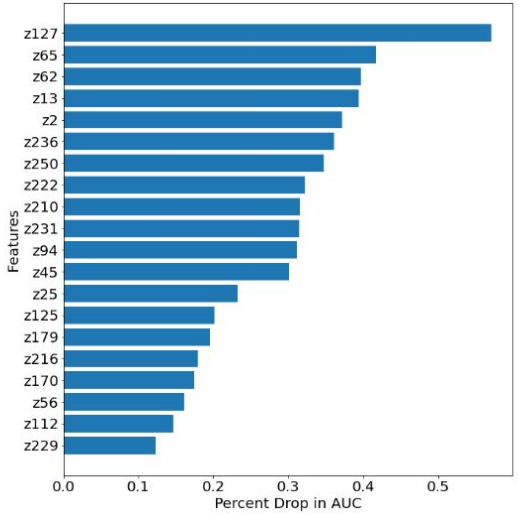
- Latent space is noticeably sparse, a compact representation is possible
- Network disentanglement before the third hidden layer in F
- Early layers of Φ are busy while F is relatively sparse
- Ample scope of model reoptimization w/o significant performance compromise

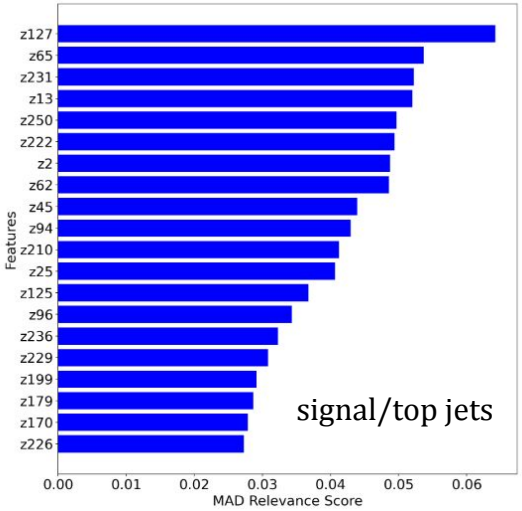


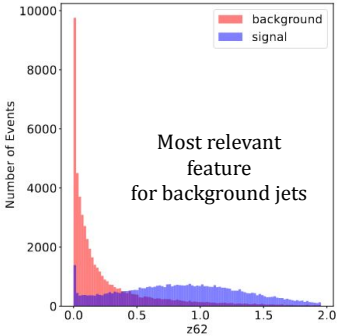
Model Architecture	ROC-AUC	Accuracy	#Parameters
Baseline	0.997	0.977	82k
Φ: (3,100,64,64), F: (64,2)	0.994	0.969	17k

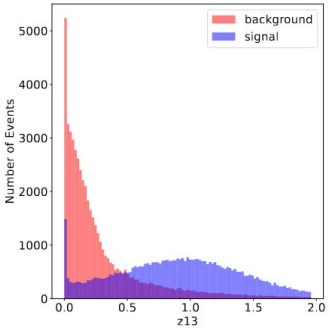
PFN's Latent Space

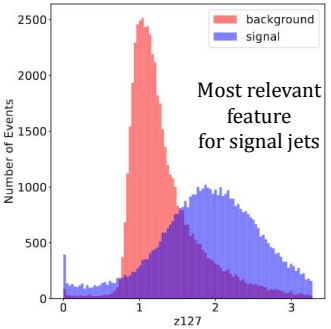
- Which latent features are important? Using occlusion test and MAD relevance scores
- Features that are "important" for different jet classes have some overlap and some non-overlapping features
- How is the network encoding the jet class information?





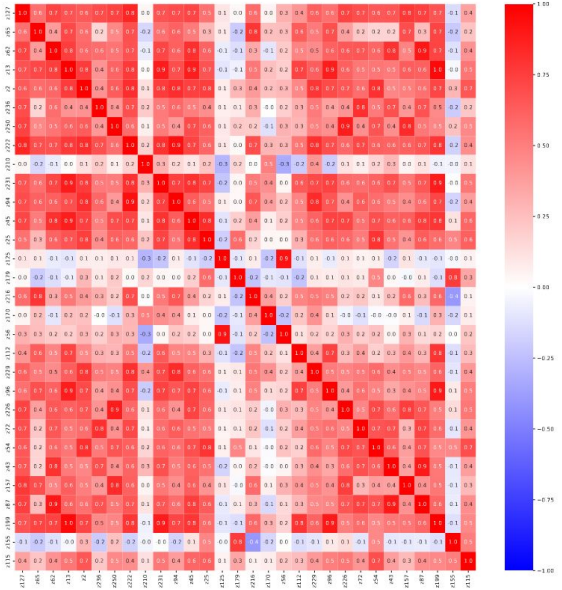




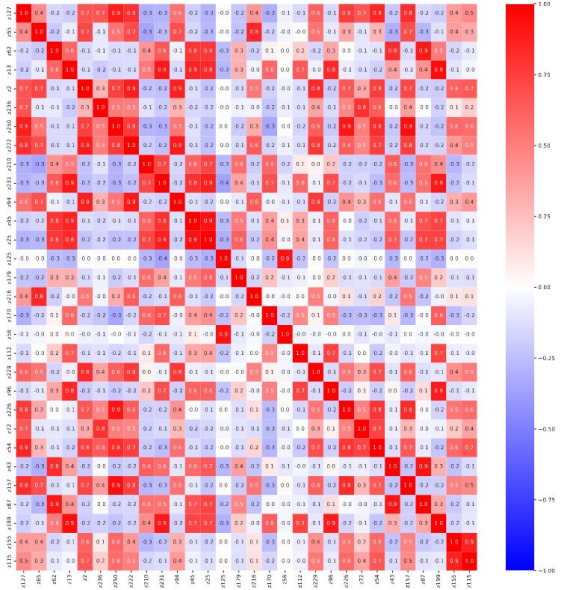


Information in Correlation

- Correlation matrices for the 30 highest ranked latent features according to occlusion test
- Very different distribution of correlations among latent dimensions
- The network encodes jet class information in correlations within the latent space



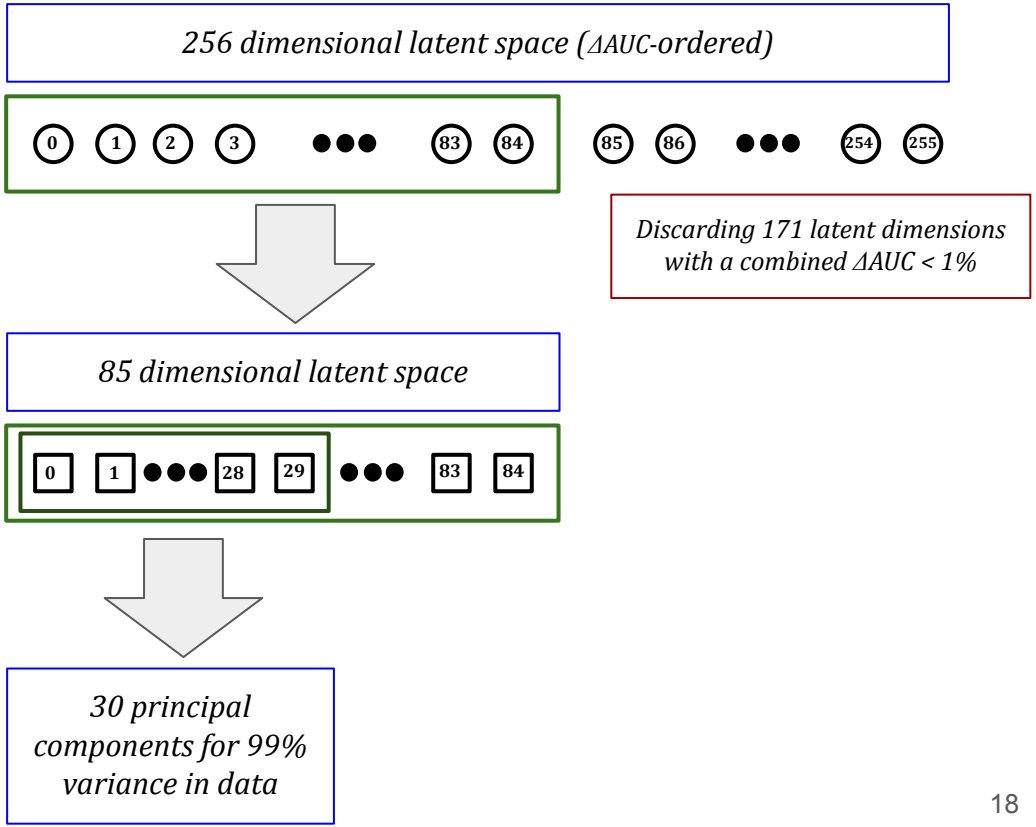
background/QCD jets



signal/top jets

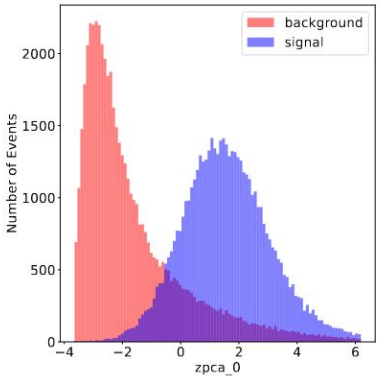
Disentangling Information from Encoded Correlations

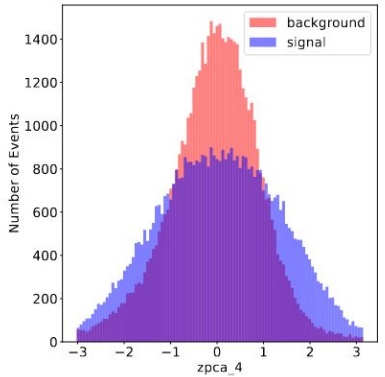
- Choose a latent subspace of highly ranked variables ($\Delta AUC < 1\%$)
- Perform Principal Component Analysis (PCA) on this latent subspace
- Select top principal components to account for up to 99% of the variance in latent data

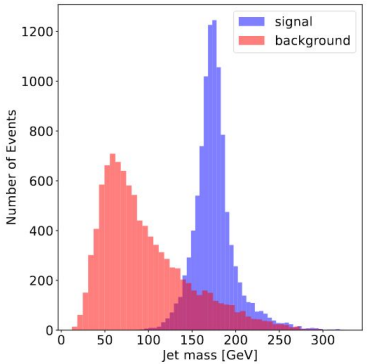


Distribution of Principal Components

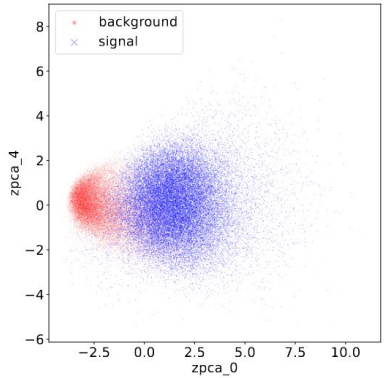
- The top principal component is an expressive feature with distinct distributions for two jet categories
- It has a very large correlation with jet mass for both jet classes (0.91 and 0.62 respectively)







The network has learned to somewhat mimic jet mass distribution w/o being trained with any such information!



Lessons Learned and Outlook

- Model interpretation can be tricky, results from existing methods can diverge
- Just like models themselves, *one size does not fit all* for model interpretation
- Be careful with model explanations (especially with LRP), especially 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