PROCEEDINGS OF SCIENCE

Interpretability of an Interaction Network for identifying $H \rightarrow b\bar{b}$ jets

Avik Roy[a,*] **and Mark S. Neubauer**[a]

[a] *Department of Physics, University of Illinois at Urbana-Champaign*
1110 W Green St Loomis Laboratory, Urbana, Illinois 61801, United States of America

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

Multivariate techniques and machine learning models have found numerous applications in High Energy Physics (HEP) research over many years. In recent times, AI models based on deep neural networks are becoming increasingly popular for many of these applications. However, neural networks are regarded as black boxes- because of their high degree of complexity it is often quite difficult to quantitatively explain the output of a neural network by establishing a tractable input-output relationship and information propagation through the deep network layers. As explainable AI (xAI) methods are becoming more popular in recent years, we explore interpretability of AI models by examining an Interaction Network (IN) model designed to identify boosted $H \rightarrow b\bar{b}$ jets amid QCD background. We explore different quantitative methods to demonstrate how the classifier network makes its decision based on the inputs and how this information can be harnessed to reoptimize the model- making it simpler yet equally effective. We additionally illustrate the activity of hidden layers within the IN model as Neural Activation Pattern (NAP) diagrams. Our experiments suggest NAP diagrams reveal important information about how information is conveyed across the hidden layers of deep model. These insights can be useful to effective model reoptimization and hyperparameter tuning.

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

[*] Speaker

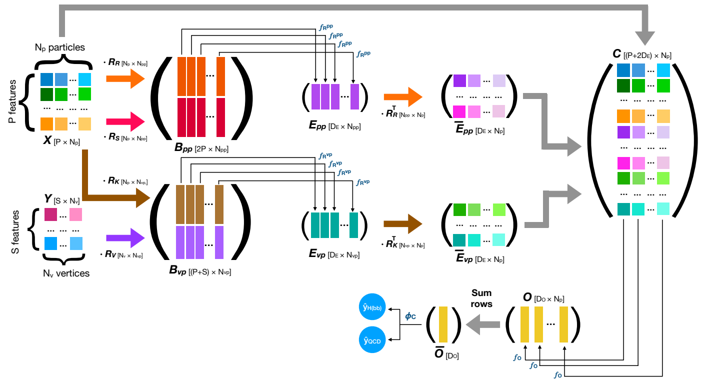
https://pos.sissa.it/

1. Introduction

Owing to their intricate internal structure, neural networks (NNs) have often been treated as *black boxes*. It is difficult to understand how different input features contribute to the network's computational process and how the inter-connected neural pathways convey information. In recent years, advances in *explainabale* Artificial Intelligence (xAI) [1] have made it possible to build intelligible relationship between an AI model's inputs, architecture, and predictions [2]. xAI has been successful in learning the underlying physics of a number of problems in high energy detectors [3]. In this work, we apply state-of-the art xAI techniques in interpreting an Interaction Network (IN) model [4] developed to identify boosted $H \to b\bar{b}$ jets from QCD background.

2. Evaluating Feature Importance for the IN Model

Figure 1 shows the IN model architecture and tabulates the default hyperparameters and data dimensions. This network is built to train on graph data structure whose nodes comprise of N_p particle tracks, each with P features, and N_v secondary vertices, each with S features, associated with the jet. The physical description of each feature is given in Appendix C of ref. [4]. It creates a fully connected directed graph with $N_{pp} = N_p(N_p - 1)$ edges for the particle tracks. A separate graph with $N_{vp} = N_v N_p$ generates all possible connections between the particle tracks and the secondary vertices.



Default IN hyperparameters	
(P, N_p, S, N_v)	(30, 60, 14, 5)
Hidden layers	3
Hidden layer dimension	60
(D_e, D_o)	(20, 24)
Activation	ReLU

Figure 1: A schematic diagram of the network architecture and dataflow in the IN model. This image is taken from Ref. [4]. The choice of model hyperparameters and input data dimensions for the baseline model is given in the accompanying table.

The node level features for the track-track (track-vertex) graph are transformed to edge level features via a couple of interaction matrices, identified as $R_{R[N_p \times N_{pp}]}$ and $R_{S[N_p \times N_{pp}]}$ ($R_{K[N_p \times N_{vp}]}$ and $R_{V[N_v \times N_{vp}]}$). These edge-level features are transformed via fully connected NNs, respectively called f_R^{pp} and f_R^{vp}, to obtain two D_e dimensional internal state representation of these graphs.

These internal state edge-level representations are transferred back to the track-level representations by the interaction matrices. These particle-level internal state representations are concatenated with the original track features creating a $(P + 2D_e) \times N_p$ dimensional feature space. The trainable dense MLP f_O creates the post-interaction D_O dimensional internal representation that are summed over the tracks and then linearly combined to produce a two-dimensional output, which is transformed to individual class probabilities via a softmax function.

In order to realize which features play the most important role in the IN's decision making process, first we train the model with its default settings (the *baseline* model). We mask one feature at a time for all input tracks or secondary vertices by replacing the corresponding entries by zero values. We obtain inference from the trained model and evaluate the Area Under the Curve (AUC) for Region Operator Characteristic (ROC) curve (ROC AUC score) from the model inference. The change observed in the AUC score for masking each of the features can be seen in Figure 2. It shows that many of these input features have a rather small impact on the model's overall performance, reflected by the very small change in AUC score. We can inspect the importance of these features for individual tracks and vertices by the Layerwise Relevance Propagation (LRP) technique [5, 6]. Since some of the input features show high degree of correlation with each other, we use the LRP-γ method described in ref. [6], which is designed to skew the LRP score distributions to nodes with positive weights in the network. In order to apply the LRP method for the IN model, we defined custom propagation functions for (i) the aggregation of internal representation and (ii) the transformation via interaction matrices.

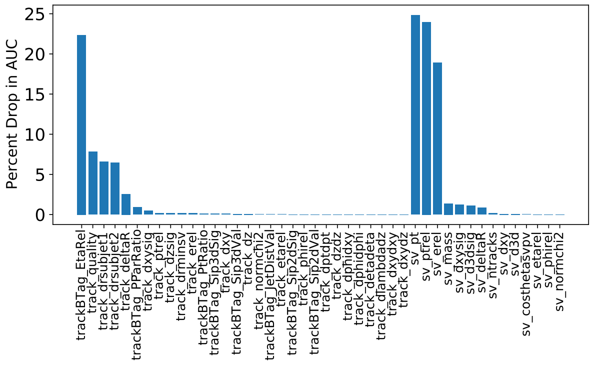
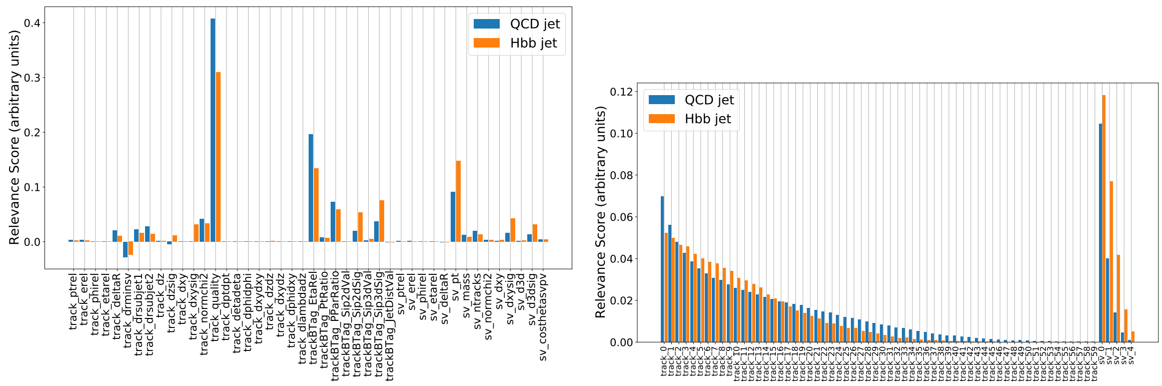


Figure 2: Change in AUC score with respect to the baseline model when each of the track and secondary vertex features is individually masked during inference with the trained baseline model.

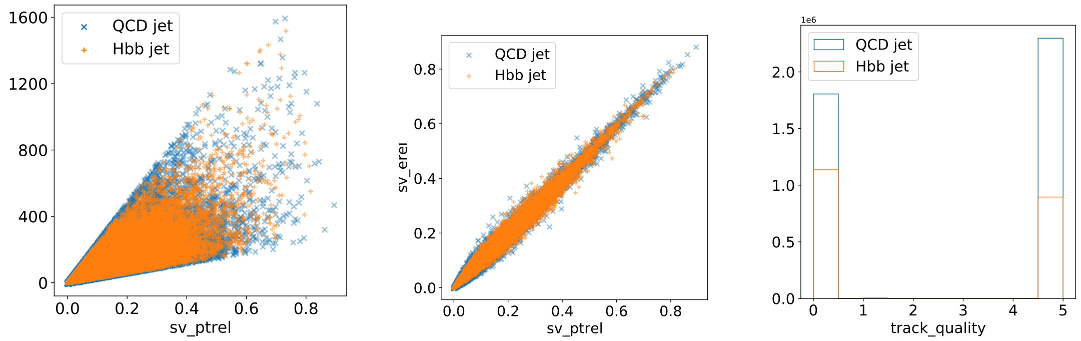
We show the average scores attributed to the different features for QCD and $H \to b\bar{b}$ jets in figure 3a. When compared with the change in AUC score by individual features in figure 2, the features with largest relevance scores coincide with the features that individually cause the largest drop in AUC score. We additionally observe that vertices features play a more important role in identifying the $H \to b\bar{b}$ jets. This behavior is also justified from a physics stand point, since the presence of high energy secondary vertices is an important signature for jets from b quark because of its relatively longer lifetime.

However, the approaches also show some inconsistencies among themselves. The secondary vertex features `sv_ptrel` and `sv_erel` are assigned very low relevance scores while masking them independently cause very large drops in the AUC score. These variables are highly correlated and both display very large correlation (correlation coefficient of 0.85) with `sv_pt` (Figures 4a and 4b). The LRP-γ method skews their relevance distribution and suppresses the LRP scores for those two variables while assigning a larger score to the variable `sv_pt`.



(a) **(b)**

Figure 3: Average relevance scores attributed to (a) input track and secondary vertex features and (b) individual tracks and secondary vertices. The tracks and secondary vertices are ordered according to their relative energy with respect to the jet energy.



(a) **(b)** **(c)**

Figure 4: Scatter plot of (a) `sv_ptrel` and `sv_pt` and (b) `sv_ptrel` and `sv_erel`

We make an additional observation regarding the importance attributed to the feature called `track_quality`. This feature is essentially a qualitative tag regarding the track reconstruction status, and for most of the training data, this has almost identical distribution for both jet categories (Figure 4c). With such an underlying distribution, it is obvious that this variable doesn't contribute to the classifier's ability to tell apart the jet categories. However, the large relevance score associated with it indicates that the classifier's class-predictive output for each class somehow receives a large contribution from its numerical embedding. The model that was trained without these variables, along with the 11 (3) track (vertex) features that report a change in AUC of less than 0.01% converged with an AUC score of 99.00%, performing as equally well.

3. Inspecting the Activation Layers and Model Reoptimization

As the IN processes the input, it is passed through three different MLPs that approximate arbitrary non-linear functions identifies as f_R^{pp}, f_R^{vp}, and f_O. In order to explore the activity of each neuron and compare it with the activity of neurons in the same layer, we define the quantity Relative Neural Activity (RNA) [7] as $\mathrm{RNA}(j,k;\mathcal{S}) = \frac{\sum_{i=1}^{N} a_{j,k}(s_i)}{\max_j \sum_{i=1}^{N} a_{j,k}(s_i)}$ where $\mathcal{S} = \{s_i\}$ represents a set of samples over which the RNA score is evaluated. The quantity $a_{j,k}(s_i)$ is the activation of j-th neuron in the k-th layer when the input to the network is s_i. Figure 5 shows the neural activation pattern (NAP) diagram for the baseline model, showing the RNA scores for the different activation layers. The scores are separately evaluated for QCD and $H \to b\bar{b}$ categories. To simultaneously visualize these scores, we project the RNA scores of the former as negative values. The NAP diagram clearly shows that the network's activity level is quite sparse- while some nodes are playing very important roles in propagating the necessary information, other nodes don't participate as much. We additionally observe that the right until the very last layer of the aggregator network f_O, the same nodes show the largest activity level for both jet categories. However, different nodes are activated in the last layer for the two jet categories, indicating an effective disentanglement of the jet category information in this layer.

The sparsity of NAP diagram and low feature importance for a number of input features for the baseline IN model indicates that the model can be made simpler, by both reducing the number of input features it relies on and the number of trainable parameters. To demonstrate this, we train some alternate variants of the IN models where we drop `track_quality`, `sv_ptrel`, `sv_erel` along with the 11 (3) track (vertex) features that report a change in AUC of less than 0.01%. The details and performance metrics of these models are given in Table 1. They demonstrate that the baseline model can be made much simpler without compromising the quality of its performance. As can be seen from the results in Table 1, the ROC-AUC score of the alternate models are very close to that of the baseline model, though the number of trainable parameters is significantly lower.

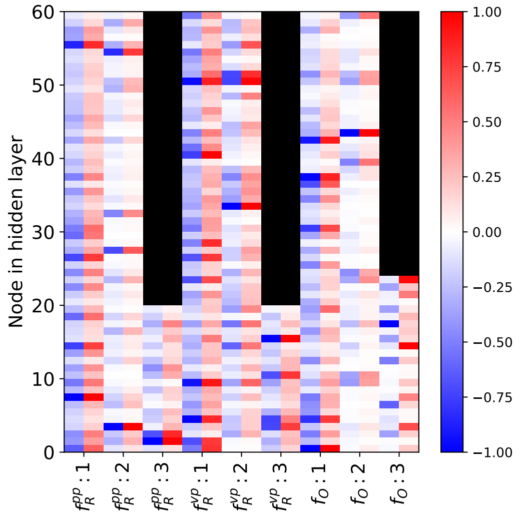


Figure 5: 2D map of RNA score for different nodes of the activation layers. To simultaneously visualize the scores for QCD and $H \to b\bar{b}$ jets, we project the RNA scores of the former as negative values

4. Conclusion

In this paper we have demonstrated how the application of xAI methods aided with physical intuitions can help identify important features for the task of identifying $H \to b\bar{b} jets$. We additionally propose a novel metric, the RNA score, and an associated visualization tool, the NAP diagram, to investigate information propagation through a model. These tools help understand the

ΔP, ΔS	h, D_e, D_o	Parameters	AUC score	Sparsity
0, 0 (baseline)	60, 20, 24	25554	99.02%	0.56
12, 5	32, 16, 16	8498	98.87%	0.52
	32, 8, 8	7178	98.84%	0.48
	16, 8, 8	2842	98.62%	0.40

Table 1: The performance of the baseline and alternate retrained models with modified hyperparameters. Sparsity is measured by the fraction of activation nodes with an RNA score less than 0.2

sparsity of information propagation and hence optimize model complexity without degrading the model's performance.

Acknowledgements

This work was supported by the U.S. Department of Energy (U.S. DOE), Office of Science, High Energy Physics, under contract number DE-SC0023365 and the FAIR Data program of the U.S. DOE, Office of Science, Advanced Scientific Computing Research, under contract number DE-SC0021258.

References

[1] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences*, *Artificial Intelligence* **267** (2019) 1.

[2] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf and G.-Z. Yang, *XAI—Explainable artificial intelligence*, *Science Robotics* **4** (2019) eaay7120.

[3] D. Turvill, L. Barnby, B. Yuan and A. Zahir, *A survey of interpretability of machine learning in accelerator-based high energy physics*, in *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, pp. 77–86, 2020, DOI.

[4] E.A. Moreno, T.Q. Nguyen, J.-R. Vlimant, O. Cerri, H.B. Newman, A. Periwal et al., *Interaction networks for the identification of boosted h→ b b decays*, *Physical Review D* **102** (2020) 012010.

[5] A. Binder, S. Bach, G. Montavon, K.-R. Müller and W. Samek, *Layer-wise relevance propagation for deep neural network architectures*, in *Information science and applications (ICISA) 2016*, pp. 913–922, Springer (2016), DOI.

[6] G. Montavon, A. Binder, S. Lapuschkin, W. Samek and K.-R. Müller, *Layer-wise relevance propagation: an overview*, *Explainable AI: interpreting, explaining and visualizing deep learning* (2019) 193.

[7] A. Khot, M.S. Neubauer and A. Roy, *A detailed study of interpretability of deep neural network based top taggers*, *arXiv preprint arXiv:2210.04371* (2022) .