Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Mixture of Experts Graph Transformer for Interpretable particle collision detection

D. Genovese, A. Sgroi, A. Devoto, S. Valentine, L. Wood, C. Sebastiani, S. Giagu, M. D'Onofrio, S. Scardapane

SAPIENZA UNIVERSITÀ DI ROMA
FAIR Future Artificial Intelligence Research
intelligent signal processing and multimedia lab
UNIVERSITY OF LIVERPOOL
EuCAIF

Explainable AI for Scientific Discovery

General Introduction

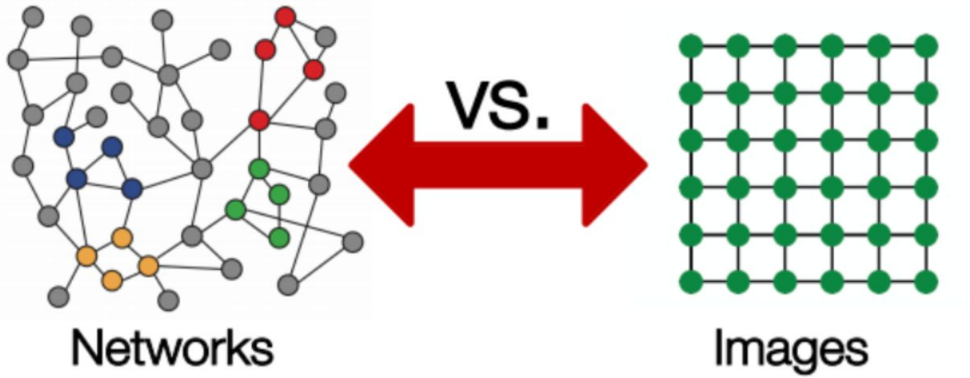
Explainable AI for Scientific Discovery

Current Techniques: Machine learning (ML) and deep learning (DL) are increasingly used in HEP data tasks for:

- Particle identification
- Event reconstruction
- Background subtraction

Graph Neural Network:

- Represents particles and interactions as nodes and edges in a graph.
- Captures complex relationships for enhanced analysis.

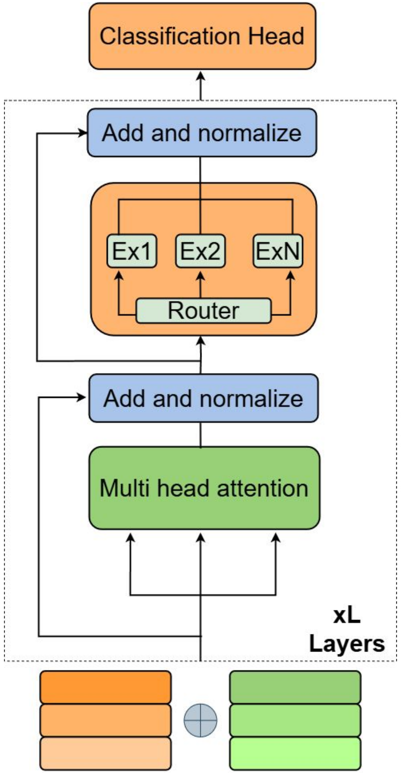


Explainable AI for Scientific Discovery

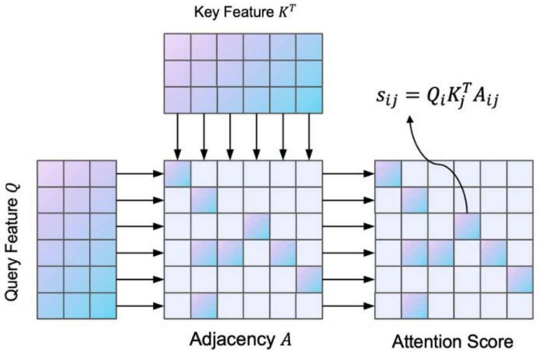
- **Interpretability Issue**: GNNs often lack transparency in decision-making processes.
- **Proposed Solution**: Introduction of a MoE Graph Transformer model.

Benefits:

- Introduce modularity with the MoE layer.
- Enhances interpretability while maintaining high predictive accuracy.
- Provides intuitive visualization of attention maps.



MoE Graph Transformer Layer



Self-Attention for Graphs

https://arxiv.org/abs/2012.09699

- Attention Mechanism in standard Transformer:

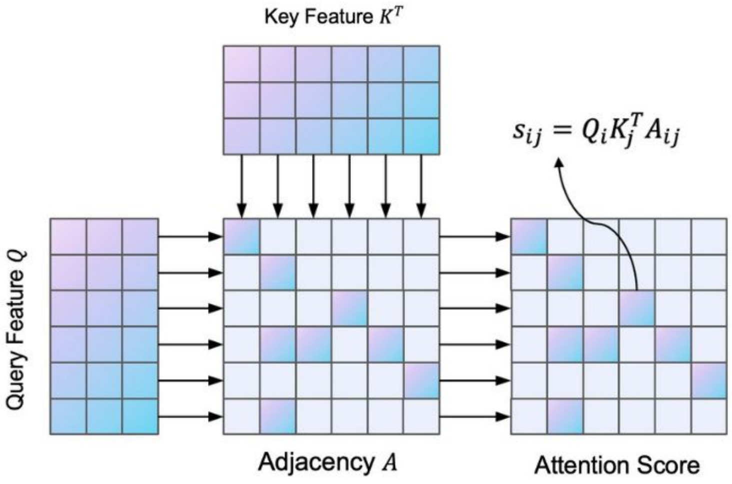
$$\mathrm{Attn} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

- Attention Mechanism for Graph Transformer:

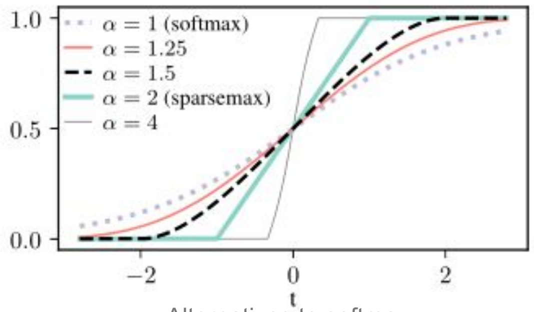
$$\mathrm{Attn}(Q, K, V, A) = \mathrm{softmax}\left(\frac{(QK^T) \circ A}{\sqrt{d}}\right)V$$

Where $Q, K, V \in \mathbb{R}^{N \times d}$ are projections of the input values and $A \in [0, 1]^{N \times N}$ is the adjacency matrix.

Only the attention scores between connected nodes are computed.



Self-Attention for Graphs



Alternatives to softmax

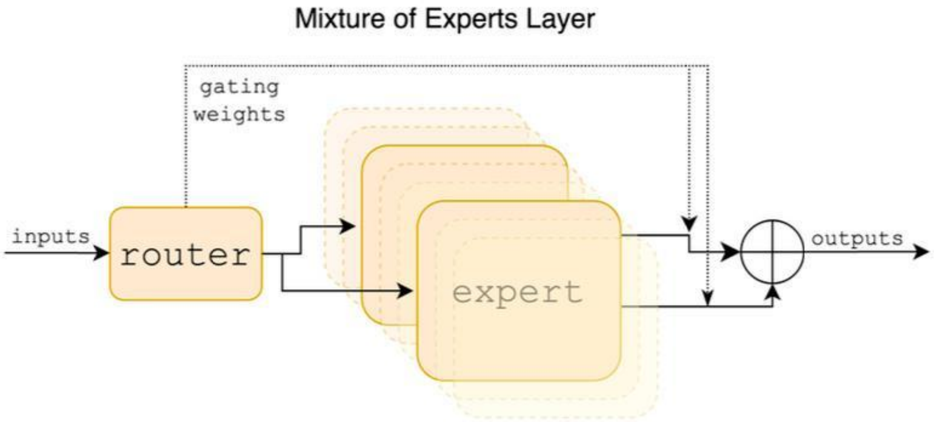
Mixture of Experts

https://arxiv.org/abs/1701.06538

- Components:
  1. One Router
  2. *N* small networks (experts)

- Layer output:

$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$



- Noisy Top-K Gating:

$$G(x) = \text{Softmax}(\text{KeepTopK}(H(x), k))$$

$$H(x)_i = (x \cdot W_g)_i + \text{StandardNormal}() \cdot \text{SoftPlus}((x \cdot W_{\text{noise}})_i)$$

MOE Graph Transformer Model

Components:

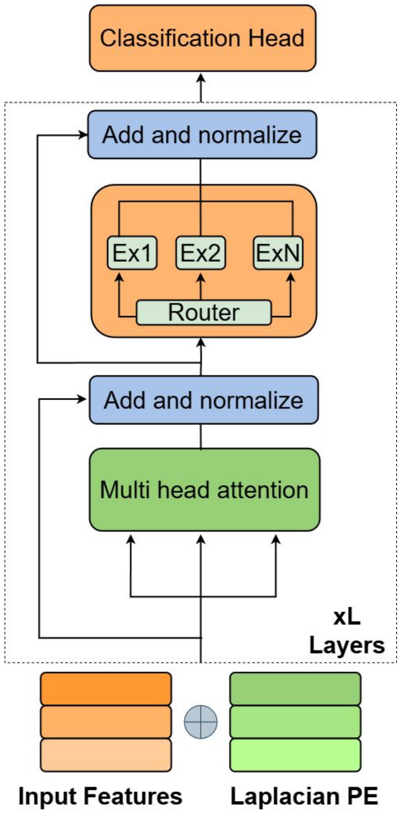
- Embedding layer
- Multi-head graph attention layer
- Layer normalization plus residual connection
- Mixture of experts layer
- Output block for classification

Load balancing Loss:

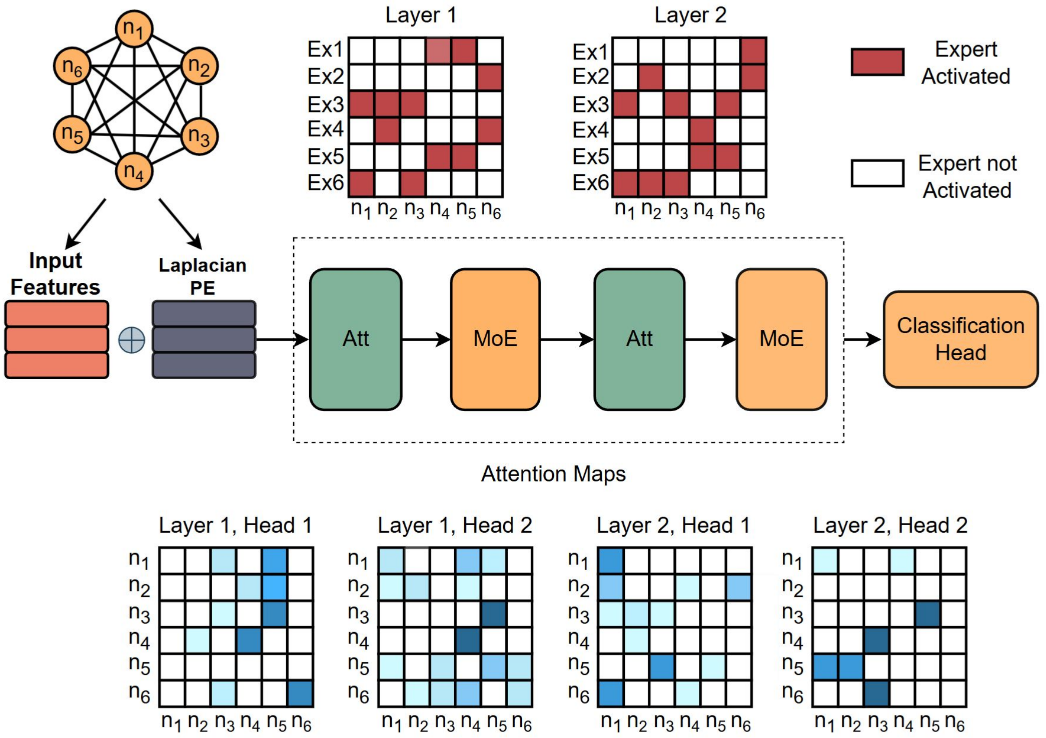
$$L_{load}(X) = w_{load} \cdot CV(Load(X))^2$$

$$Load(X)_i = \sum_{x \in X} P(x, i)$$

$$P(x, i) = \Phi\left(\frac{(x \cdot W_g)_i - kth\_excluding(H(x), k, i)}{Softplus((x \cdot W_{noise})_i)}\right)$$



Architecture



Overview of the architecture

The graph is processed by the model components:

- Multi-Head Attention block
- MoE block
- Classification head

The visualization includes attention maps derived from the MultiHead Attention mechanism and the activation patterns of the experts for a single collision event example.

The Dataset

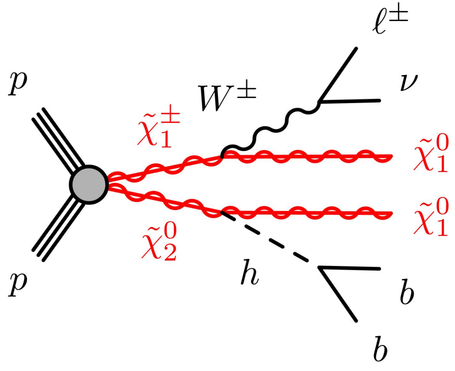
The Dataset

The Dataset is available at: https://opendata.cern.ch/record/28100

Dataset Characteristics

Three types of graphs:

- Signal
- ttbar (background)
- Single-top (background)



Particle	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
jet1	'pTj1'	'etaj1'	'phij1'	'j1_quantile'	nan	nan
jet2	'pTj2'	'etaj2'	'phij2'	'j2_quantile'	nan	nan
jet3 (optional)	'pTj3'	'etaj3'	'phij3'	'j3_quantile'	nan	nan
b1	'pTb1'	'etab1'	'phib1'	'b1_quantile'	'b1m'	nan
b2	'pTb2'	'etab2'	'phib2'	'b2_quantile'	'b2m'	nan
lepton	'pTl1'	'etal1'	'phil1'	nan	nan	nan
energy	'ETMiss'	nan	'ETMissPhi'	nan	nan	'metsig_New'

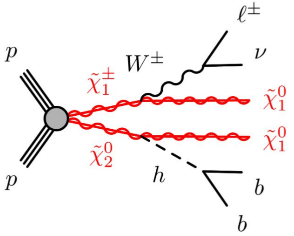
Features of nodes

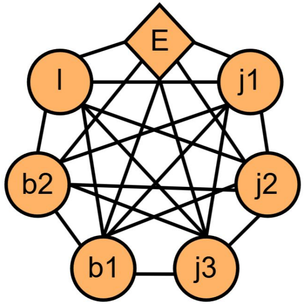
The Dataset

The Dataset is available at: https://opendata.cern.ch/record/28100

Dataset Characteristics

- Each row of the dataset contains 1 graph of 6 or 7 nodes.
- Each graph is fully connected.
- Each graph has a maximum of 6 features.





Representation of a collision event and the corresponding graph

Particle	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
jet1	'pTj1'	'etaj1'	'phij1'	'j1_quantile'	nan	nan
jet2	'pTj2'	'etaj2'	'phij2'	'j2_quantile'	nan	nan
jet3 (optional)	'pTj3'	'etaj3'	'phij3'	'j3_quantile'	nan	nan
b1	'pTb1'	'etab1'	'phib1'	'b1_quantile'	'b1m'	nan
b2	'pTb2'	'etab2'	'phib2'	'b2_quantile'	'b2m'	nan
lepton	'pTl1'	'etal1'	'phil1'	nan	nan	nan
energy	'ETMiss'	nan	'ETMissPhi'	nan	nan	'metsig_New'

Features of nodes

The Dataset

The Dataset is available at: https://opendata.cern.ch/record/28100

Dataset Characteristics

Three types of graphs:

- Signal
- ttbar (background)
- Single-top (background)

Task

Training a model that performs binary classification (i.e. recognizes signal and background events)

Particle	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
jet1	'pTj1'	'etaj1'	'phij1'	'j1_quantile'	nan	nan
jet2	'pTj2'	'etaj2'	'phij2'	'j2_quantile'	nan	nan
jet3 (optional)	'pTj3'	'etaj3'	'phij3'	'j3_quantile'	nan	nan
b1	'pTb1'	'etab1'	'phib1'	'b1_quantile'	'b1m'	nan
b2	'pTb2'	'etab2'	'phib2'	'b2_quantile'	'b2m'	nan
lepton	'pTl1'	'etal1'	'phil1'	nan	nan	nan
energy	'ETMiss'	nan	'ETMissPhi'	nan	nan	'metsig_New'

Features of nodes

Training-Susy Dataset

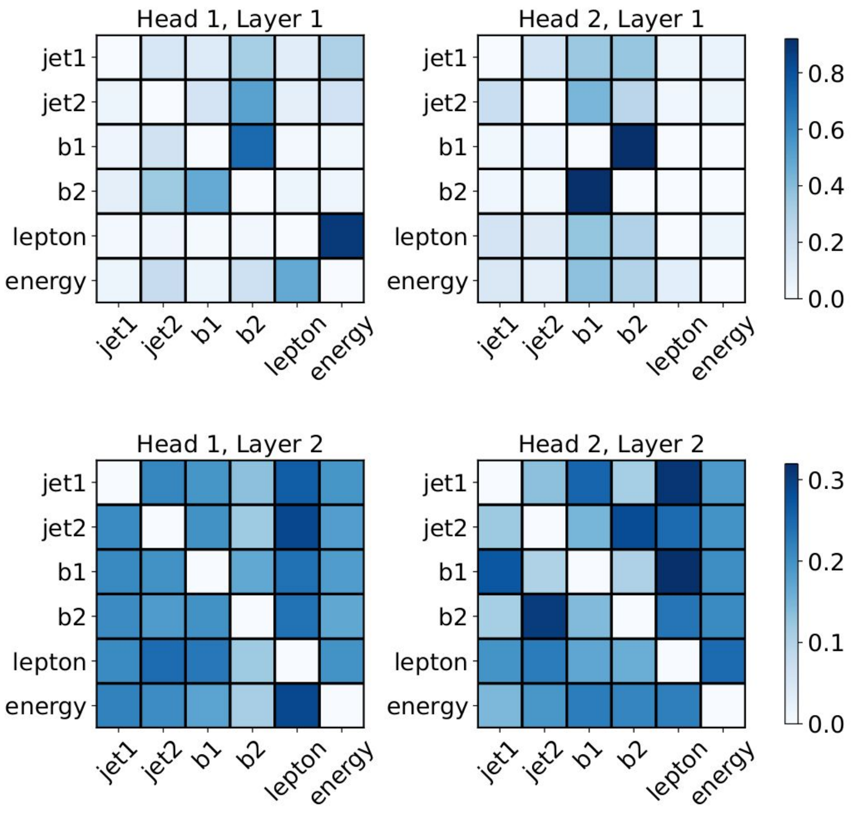
Model and Hyperparameters

- MoE Graph Transformer with 2 layer and 2 heads
- Adam optimizer with a learning rate of 2e^(-3)
- 60 epochs of training
- Cross entropy loss
- Batch size of 512
- Signal events: 400k
- ttbar events: 200k
- Single-top events: 200k

Test results

Model	Accuracy	AUC
GCN	0.750 +/- 0.0022	0.832 +/- 0.134
MLP	0.829 +/- 0.0015	0.913 +/- 0.0017
GT	0.849 +/- 0.0059	0.928 +/- 0.0057
MGT	0.852 +/- 0.0005	0.929 +/- 0.0039

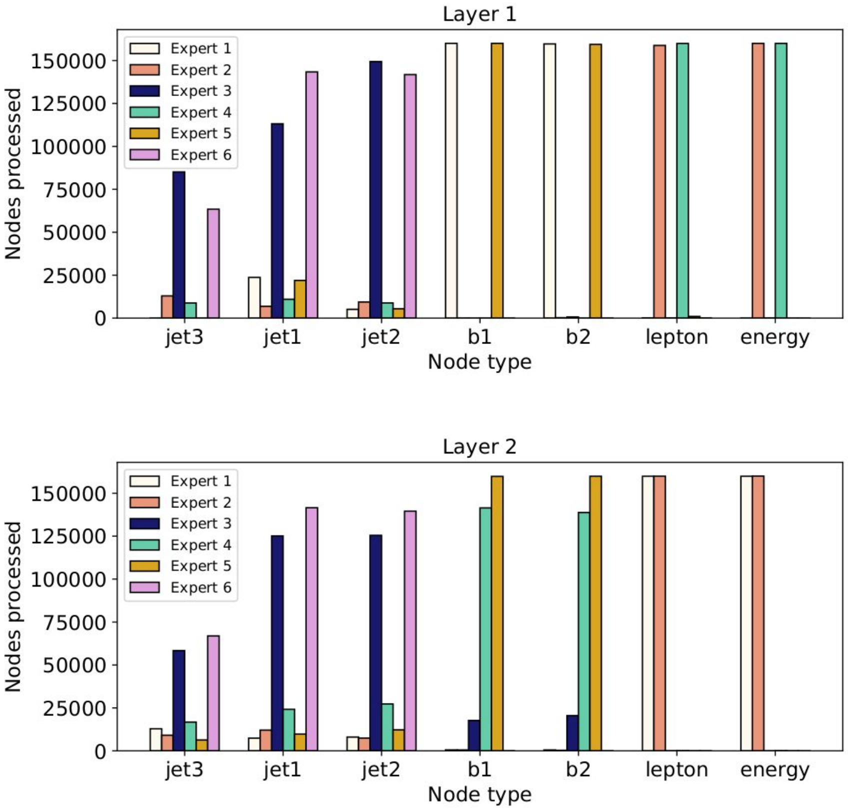
Explainability results: Attention maps for test set



First layer: displays higher magnitudes and broader distributions.
Second layer: shows reduced magnitudes, indicating a refinement of features.

The attention focus is primarily on core features of b1,b2 for head 2 and of energy, lepton for head 1

Explainability results: Experts specialization



Expert Number	Specialization	Physical Explanation
1, 5	*b1*, *b2*	Experts focus on the b-jets as a correlated entity. In signal events, these arise from Higgs boson decay, but in background events, they do not.
2, 4	*lepton*, *energy*	Experts specialize in detecting lepton and missing energy features, key indicators for distinguishing signal from background. In signal, missing energy also originates from dark matter particles, unlike the background.
3, 6	*jet1*, *jet2*, *jet3*	Equivalent to *b1* and *b2* in 6-node graphs, but in 7-node graphs, the additional jet (*jet3*) introduces distinct kinematic features that alter the specialization

Test on different configurations

Signal models were parameterized by the mass of the SUSY parent particles and the mass of the dark matter particle. The mass value ranges between 125 GeV and 1000 GeV for the former, and 0 and 400 GeV for the latter mass.

Model	Accuracy	AUC
Group 1	0.683+/-0.0165	0.756+/-0.021
Group 2	0.748+/-0.019	0.835+/-0.012
Group 3	0.923+/-0.009	0.977+/-0.003
Mixed	0.852 +/-0.0005	0.929 +/-0.0039

Different configurations of data

Model	Mass difference
Group 1	150 GeV
Group 2	200 GeV
Group 3	>500 GeV

Future improvements

- Test the model on **different datasets** to verify its applicability.
- Evaluate its ability to **generalize across diverse high-energy physics tasks.**
- Integrate **Large Language Models** for automated, accessible explanations

Thanks for the attention