



PAPER • OPEN ACCESS

Graph transformer and mixture of experts for rare signal detection in calorimeter data

To cite this article: Donatella Genovese *et al* 2026 *Mach. Learn.: Sci. Technol.* **7** 045049

View the [article online](#) for updates and enhancements.

You may also like

- [Graph neural networks in particle physics](#)
Jonathan Shlomi, Peter Battaglia and Jean-Roch Vlimant
- [Partition pooling for convolutional graph network applications in particle physics](#)
M. Bachlechner, T. Birkenfeld, P. Soldin et al.
- [Supervised Deep Learning in High Energy Phenomenology: a Mini Review](#)
Murat Abdughani, , Jie Ren et al.



PAPER

OPEN ACCESS

RECEIVED
11 May 2026REVISED
17 July 2026ACCEPTED FOR PUBLICATION
21 July 2026PUBLISHED
12 August 2026

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Graph transformer and mixture of experts for rare signal detection in calorimeter data

Donatella Genovese^{1,*} , Alessio Devoto¹ , Joseph Carmignani² , Cristiano Sebastiani³ ,
Simone Scardapane⁴ and Monica D'Onofrio²

¹ Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy

² Department of Physics, University of Liverpool, Liverpool, United Kingdom

³ Department of Experimental Physics, CERN, Geneva, Switzerland

⁴ Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, Rome, Italy

* Author to whom any correspondence should be addressed.

E-mail: donatella.genovese@uniroma1.it

Keywords: dark photons, graph neural networks, high-energy physics, machine learning

Supplementary material for this article is available [online](#)

Abstract

This work investigates graph-based deep learning approaches for the identification of rare and non-standard particle signatures in the ATLAS experiment at CERN focusing on dark photons, new particles predicted by beyond standard model theories, as a representative case. Such signatures arise from displaced decays and produce highly heterogeneous and irregular energy deposition patterns in part of the ATLAS detector such as the calorimeter, that are difficult to model with conventional approaches. Graph neural networks provide a natural framework to model the sparse and relational structure of these data. We evaluate a range of architectures, including message-passing networks, attention-based models, and graph Transformers. Building on this, we introduce a graph Transformer augmented with a mixture of experts (MoE) mechanism, designed to better capture the intrinsic heterogeneity of these signatures through modular representations. The models are benchmarked on public, simulated ATLAS calorimeter data, and their behaviour is analysed using physics-motivated observables. The results show that graph-based approaches, and in particular the MoE-enhanced graph Transformer, improve sensitivity to rare and complex signal topologies, while providing insight into the learned representations, and as such offer useful guidance for future applications in particle physics and related domains.

1. Introduction

The rapid advancement of machine learning has profoundly impacted various scientific fields, with high-energy physics (HEP) being no exception, as the field increasingly requires sophisticated tools to analyse complex datasets [1]. In this context, the large Hadron Collider (LHC) at CERN provides a unique environment for studying the fundamental constituents of matter through high-energy proton–proton collisions. Experiments such as ATLAS [2] record vast amounts of data from these interactions, enabling detailed investigations of particle properties and rare processes across a wide energy range. Advanced machine learning techniques are increasingly required to extract subtle and complex signatures from the detector data.

In recent years, deep learning methods, particularly convolutional neural networks (CNNs) [3], have played a pivotal role in tasks such as jet tagging, particle identification, and event classification in experiments at the LHC [4]. CNNs exploit the structured nature of image-like data, making them well-suited for inputs represented on regular grids. Nonetheless, data in HEP often exhibit irregular, sparse, and non-Euclidean structures that are not easily captured by standard convolutional operations. For example, the energy deposits in ATLAS calorimeter cells [5] or the spatial arrangement of detector hits naturally form point clouds or graphs, where the relationships between elements are more complex than simple

pixel adjacency. In such cases, CNNs can struggle to effectively exploit the underlying geometry and connectivity of the data without significant preprocessing or loss of information.

CNNs limitations are particularly relevant for tasks involving rare or non-standard signatures, where subtle spatial correlations and irregular topologies carry critical information. To address those, graph neural networks (GNNs) [6] have emerged as a powerful alternative, offering a framework capable of learning directly from graph-structured data. In this context, nodes can represent reconstructed calorimeter clusters or detector hits, while edges encode spatial or physical relationships between them. These models offer the ability to capture intricate relationships between calorimeter clusters, which are essential for identifying new particles [7]. More recently, approaches based on particle-cloud representations have further improved performance in collider physics tasks. In particular, ParticleNet [8] models jets as sets of particles and applies dynamic graph convolutional networks to learn local relationships between neighbouring particles. This approach demonstrated that directly operating on particle clouds can significantly improve jet tagging performance compared to image-based methods.

Recent advances have also introduced transformer-based architectures for modelling particle collision data, several of which have been successfully applied to jet tagging tasks, including recent developments within the ATLAS experiment where transformer-based and graph-based models are employed for jet flavour tagging [9]. In particular, the particle Transformer (ParT) [10] represents individual particles as token embeddings and applies self-attention mechanisms to learn interactions between all particles in an event. By modelling global relationships through attention, ParT is able to capture complex correlations across the entire particle set, achieving state-of-the-art performance on jet tagging benchmarks. A related approach is adopted by the OmniJet- α architecture [11], which first tokenises the continuous kinematic features of reconstructed particles into discrete representations and then processes them with a Transformer model.

Despite advances on both GNN and Transformer, the application of graph Transformer (GT) architectures [12] to calorimeter-based representations remains relatively unexplored. Calorimeter data consist of sparse and irregular energy deposits across detector cells, where the geometric relationships between clusters carry essential information about the underlying particle interactions. Moreover, these data are inherently heterogeneous, as different physical processes can produce widely varying spatial and energetic patterns. Effectively modelling these relationships is therefore crucial for identifying rare physics signatures.

This work contributes to the development of scalable and interpretable machine-learning frameworks for complex detector data in large-scale experiments such as ATLAS. In this work, we focus on the identification of dark photons in the ATLAS detector, which we adopt as a representative case of rare and non-standard signatures at colliders [13]. The task is to distinguish displaced dark-photon decays in the calorimeter from Standard Model background and from rare non-standard detector signatures, using cluster-level information that captures the fine spatial and energetic structure of the shower. This is particularly challenging because the signal can be highly heterogeneous, depending on the dark-photon mass, lifetime, and decay position, and because its calorimeter signature may differ significantly from that of prompt photons and jets while still being affected by detector granularity, noise, and pile-up. Building upon previous applications of deep learning in HEP, we propose a GT extended with a mixture of experts (MoE) mechanism [14] as the main architecture for dark photon identification. The MoE module replaces the standard feed-forward sub-layer in each Transformer block with a set of specialised expert networks, enabling a modular decomposition of the feature space while increasing model capacity. This design is particularly suited to handling heterogeneous and irregular topologies, as different experts can adapt to distinct classes of calorimeter signatures. To systematically assess the benefit of this architecture, we benchmark it against established GNN methods (graph convolutional network (GCN), graph attention network (GAT) [15]) and a CNN baseline operating on image representations of the same data, contributing to ongoing efforts to probe beyond the standard model (BSM) at the LHC.

2. Dark photon use case

We focus on a specific HEP application of particle classification in simulated data with the ATLAS calorimeter detectors. The goal is to investigate the potential of machine learning techniques for classifying energy deposition patterns associated with different particles, with particular emphasis on rare and non-standard signatures. The study utilises Monte Carlo simulations of the ATLAS detector, which are publicly available from previous ATLAS papers [16]. These simulations provide detailed representations of energy deposits in the calorimeters from both standard model (SM) particles, as well as hypothetical new particles known as dark photons. The signal model considers the decay of a Higgs boson into

two dark photons. The dark photon is a massive gauge boson associated with a hidden $U(1)$ sector that can decay into leptons and light hadrons through kinetic mixing with the SM photon. Its proper lifetime depends on the kinetic-mixing strength, with smaller values leading to increasingly displaced decays. As a consequence, the resulting detector signature consists of a narrow jet that originates inside the calorimeter rather than from the interaction point, producing the displaced calorimeter showers targeted in this study. The simulations account for realistic detector effects, including resolution limitations and noise, making the task of distinguishing new physics signals from the SM background particularly challenging. The primary goal is to discriminate rare new physics events, which appear as unconventional late energy deposits, from a dominant background of known particles that can mimic these signals. This setting provides a representative benchmark for studying machine learning methods in the presence of heterogeneous and irregular detector signatures. Our investigation aims to develop an ML model capable of accurately identifying these rare events and also to provide clear and interpretable explanations for the model's predictions, linking them to physically meaningful observables.

3. Dataset

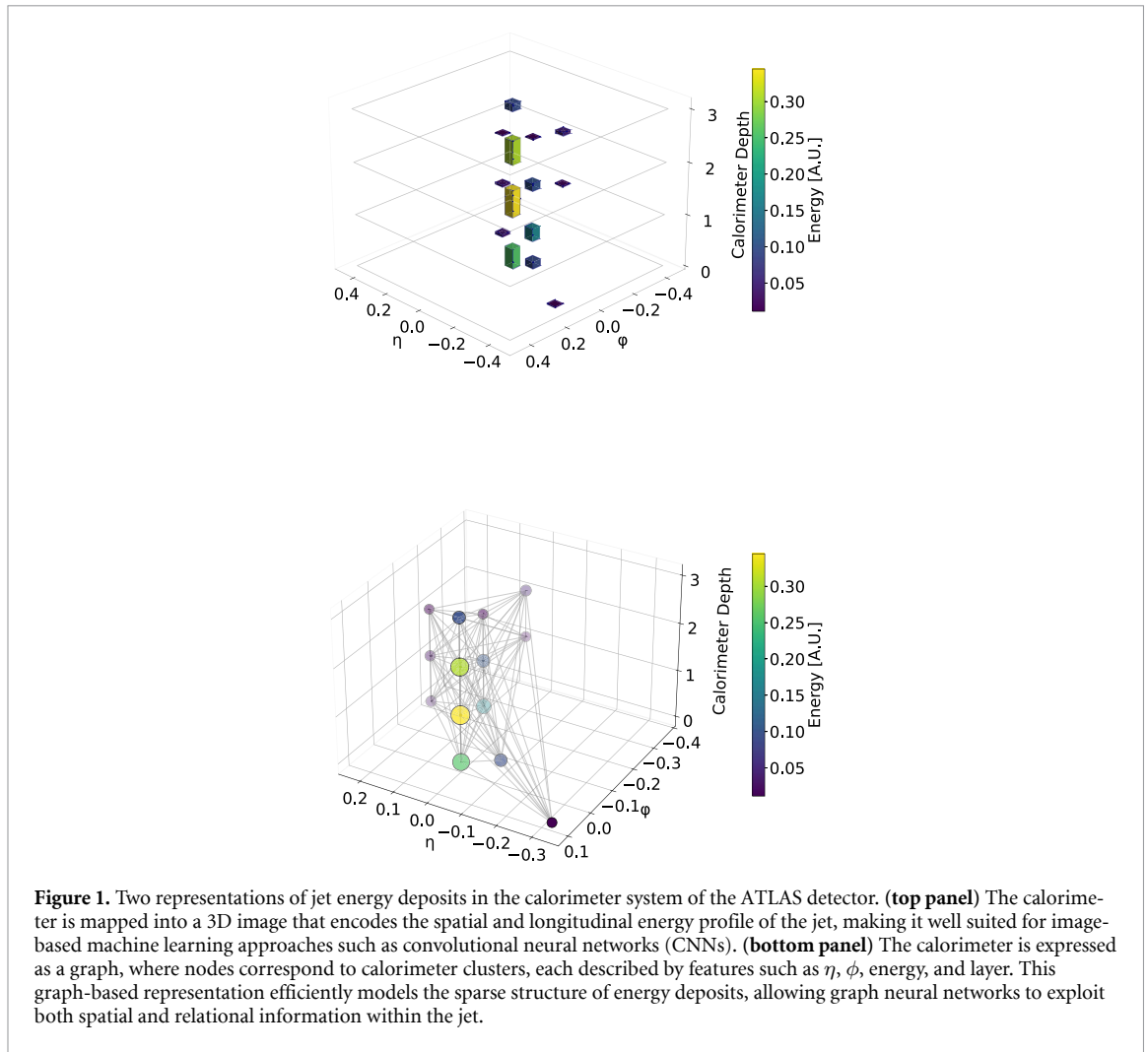
The simulated dataset employed in this work corresponds to the publicly available ATLAS benchmark released together with the original dark-photon search through HEPData [16]. For completeness, we briefly summarise below the main aspects of the event generation and reconstruction.

The signal sample assumes the production of dark photons through the decay of a SM Higgs boson with a mass of 125 GeV. The matrix-element calculation was performed at leading order. The dominant SM background consists of standard jets from quarks and gluons, generated using the same tuned parameter set and parton distribution functions as the signal samples. The generated events were processed through a detailed simulation of the ATLAS detector [17] based on Geant4 [18]. The simulation includes the effects of multiple pp interactions occurring in the same bunch crossing as the hard-scattering interaction, as well as in neighbouring bunch crossings (pile-up), thereby providing a realistic description of detector conditions. At reconstruction level jets are required to have an electromagnetic energy fraction, defined as the ratio of the energy deposited in the electromagnetic calorimeter to the total jet energy, smaller than 0.4. Reconstructed jets are required to satisfy $p_T > 20$ GeV and to lie within the pseudorapidity range $|\eta| < 2.5$.

The ATLAS calorimeter system plays a central role in measuring the energy released by particles produced in proton–proton collisions. When high-energy quarks or gluons are created, they cannot exist freely due to colour confinement and instead hadronise into collimated sprays of particles known as jets. These jets leave clustered energy deposits in the calorimeter, both in the electromagnetic and hadronic sections, which are then aggregated across multiple sampling layers to reconstruct the jet. The reconstruction process provides key jet observables [2] such as pseudorapidity (η), azimuthal angle (ϕ), and transverse momentum (p_T), which are essential to characterise the kinematics and dynamics of the underlying hard-scattering process. The resulting detector data are inherently sparse, irregular, and heterogeneous, reflecting the complexity of the underlying particle interactions. Energy deposits from jets can be represented in two primary formats for this analysis. One approach maps the deposits onto a fixed 3D grid (see figure 1, top panel), reflecting the calorimeter's layered structure. These 3D grids are treated as images, making them suitable inputs for CNNs. Each image is built from a 15×15 grid, centred on the jet axis and covering a region of $\eta \times \phi = 0.9 \times 0.9$. Each cell corresponds to a 0.06×0.06 area in the $\eta \times \phi$ plane. A third axis represents the calorimeter sampling layers, creating a 3D array where each pixel contains the energy deposited at a specific spatial location and depth. This structure captures the spatial distribution of energy in relation to the jet, providing a rich input for image-based analysis.

An alternative representation models the same calorimeter data as graphs, where each jet corresponds to one graph, as depicted in figure 1, bottom panel. The nodes in the graph represent individual calorimetric clusters and are described by features such as η , ϕ , energy and the sampling layer. Edges are defined based on the geometric distance between nodes in the $\eta \times \phi$ space, with the edge attribute capturing this distance. This graph-based representation is particularly advantageous for processing sparse data, as it allows the network to exploit both the spatial distribution and the relational structure of energy deposits within a jet, while preserving the irregular geometry of the detector response.

We preprocess our data by filtering out all the events for which all energy deposits are located in a single calorimeter layer. Although such events may occur in real detector data, they contain little or no longitudinal shower information, effectively reducing the problem to a two-dimensional pattern. Since both the CNN and graph-based representations are specifically designed to exploit the full



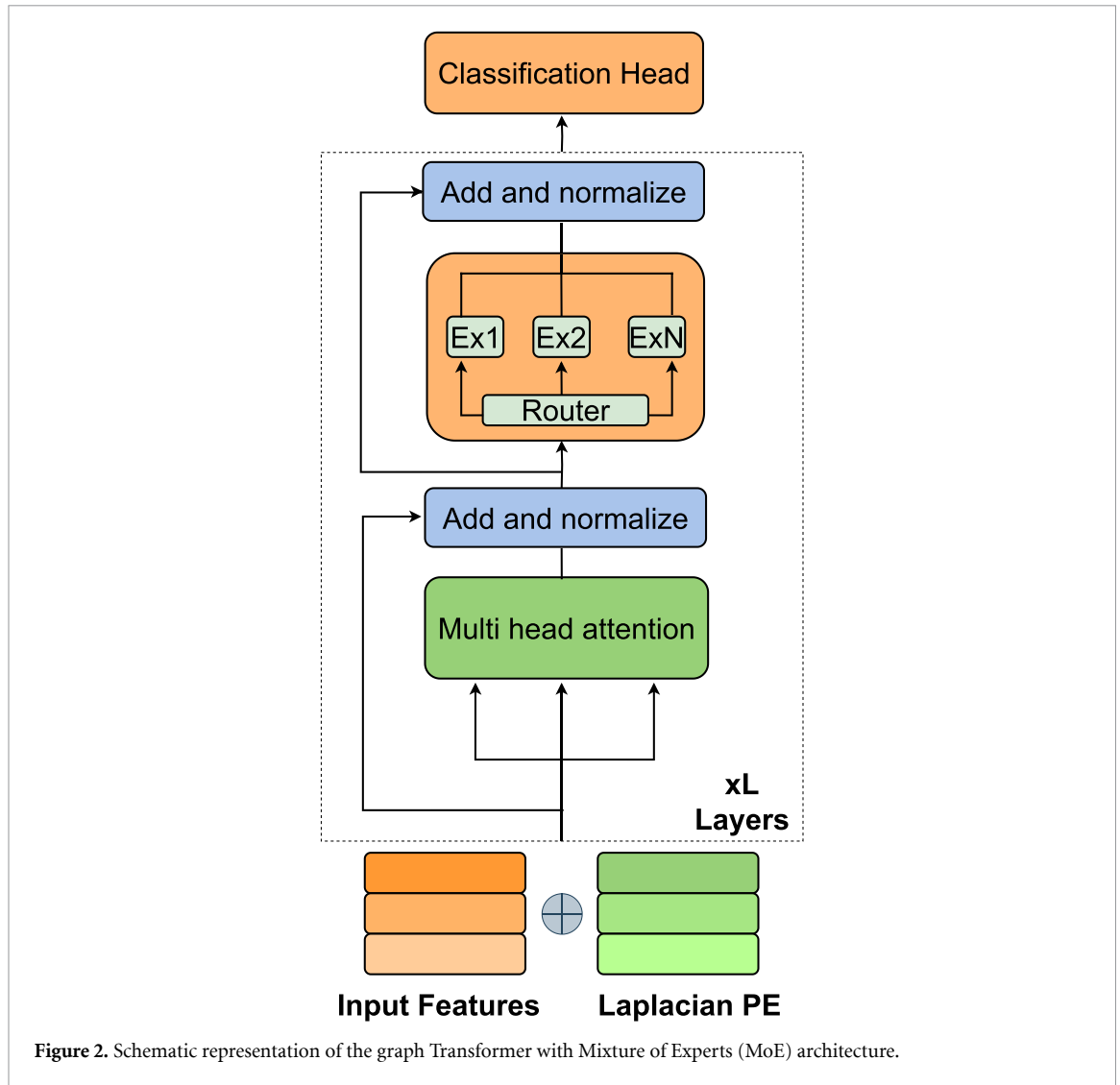
three-dimensional calorimeter structure, these events provide limited discriminating information for the present study. We further verified that this selection does not affect the generalisation capability of the model: a classifier trained without these events and evaluated on a test set that includes them exhibits no measurable degradation in overall performance. Retaining these events would therefore provide limited physical information while potentially biasing the training towards topologies that are not representative of the three-dimensional shower development targeted by the proposed architectures.

4. Methods

We adopt as our primary model a **GT with MoE** [19], specifically designed to exploit both the relational structure of the calorimeter data and the heterogeneity of jet topologies. Each jet is represented as a graph with node features of dimension 4, corresponding to the cluster kinematic variables and sampling layers. The model consists of 2 Transformer encoder layers, each equipped with 2 attention heads and ReLU activations in the feed-forward blocks. Positional information is injected through two-dimensional Laplacian eigenvectors [20] (Laplacian PE), which provide a spectral encoding of the graph structure and allow the attention mechanism to exploit the underlying geometry of the calorimeter clusters.

The standard feed-forward sub-layer in each Transformer block is replaced by a MoEs module with 6 experts, each implemented as a feed-forward network. A top- $k = 2$ sparse gating strategy is employed so that, for each input, only the two most relevant experts are activated, and a load-balancing loss term is added to encourage a uniform utilisation of the experts across the dataset. This design increases the effective model capacity, which is expected to promote specialisation of different experts to distinct regions of the feature space, hence is particularly beneficial for capturing rare signal-like patterns within a large background. A schematic illustration of the architecture can be seen in figure 2.

To assess the benefit of this architecture, we compare it against different graph reference models. As a first graph-based baseline, the **GCN** [21] employs two graph convolution layers with ReLU activations,



followed by batch normalisation and dropout. The hidden dimensionality is kept fixed across layers, and node embeddings are aggregated into a graph-level representation using global mean pooling before the final classification layer. This architecture captures local neighborhood information via message passing but does not include attention-based weighting.

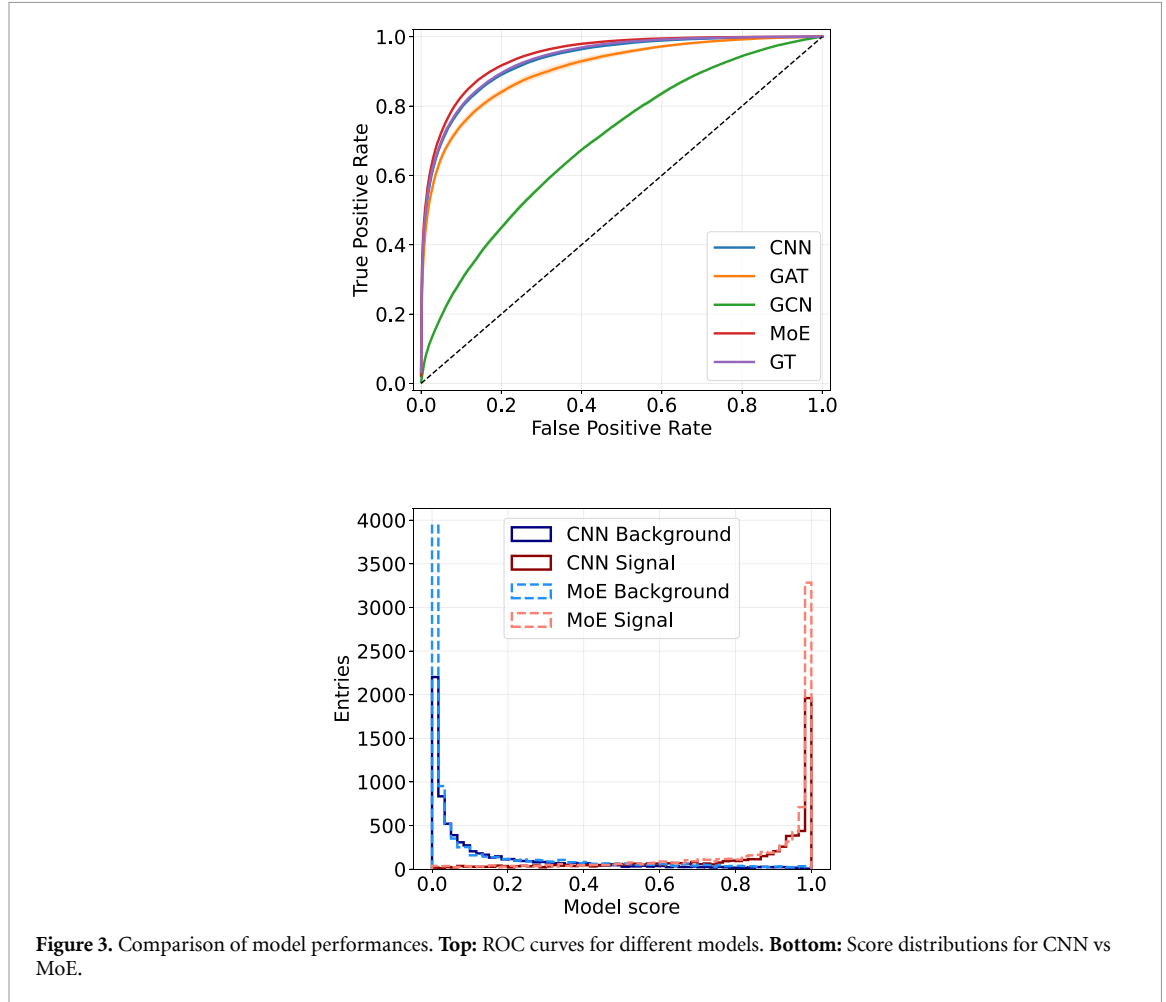
The second graph-based baseline is the **GAT** [15], which replaces uniform neighborhood aggregation with learned attention coefficients. It includes two attention layers: the first with 4 attention heads and ELU activation, and the second with a single head and ELU. Both layers are followed by batch normalisation and dropout, and a global mean pooling operation is used to obtain a graph-level representation for classification.

Finally, we include a **GT** baseline to disentangle the impact of the Transformer architecture from that of the MoEs. The GT model follows the same overall Transformer design as the MoE variant, with 2 Transformer encoder layers and 2 attention heads per layer. Node features of dimension 4 are projected to a hidden size of 40, and positional information is encoded using two-dimensional Laplacian eigenvectors. Each encoder layer includes a standard feed-forward network with hidden dimension 35 and ReLU activation, without expert routing or sparse gating. A dropout rate of 0.2 is applied within the encoder layers. This baseline allows for a direct comparison between dense and expert-based feed-forward transformations within a GT framework.

In addition to the graph-based models, we include a **CNN** baseline operating on image representations of the calorimeter data, following the approach adopted in the previous ATLAS studies of dark photon identification [13]. In this setting, energy deposits are mapped onto fixed three-dimensional grids, allowing the use of convolutional filters to capture local spatial patterns across calorimeter layers. This baseline provides a reference for comparison with established image-based techniques and enables a direct assessment of the advantages of graph-based representations for modelling irregular detector data.

Table 1. Performance comparison of different neural network architectures for jet classification on the ATLAS dataset.

Model	Accuracy	F1-score	Recall	AUROC
GCN	0.639 ± 0.004	0.585 ± 0.016	0.542 ± 0.004	0.682 ± 0.009
GAT	0.812 ± 0.010	0.768 ± 0.018	0.665 ± 0.034	0.906 ± 0.006
MoE	0.867 ± 0.002	0.857 ± 0.003	0.852 ± 0.007	0.943 ± 0.002
CNN	0.850 ± 0.003	0.839 ± 0.005	0.824 ± 0.025	0.931 ± 0.002
GT	0.854 ± 0.001	0.858 ± 0.007	0.827 ± 0.010	0.934 ± 0.001

**Figure 3.** Comparison of model performances. **Top:** ROC curves for different models. **Bottom:** Score distributions for CNN vs MoE.

5. Results and performances

The classification results for all architectures are summarised in table 1 and in figure 3. Figure 3(top panel) shows the ROC curves for all considered architectures, providing a global comparison of their discriminative performance. The bottom panel shows the score distributions for the CNN and MoE models specifically. These represent the two most contrasting paradigms compared in this work, with the CNN operating on image-based representations of the calorimeter data, and the MoE exploiting the graph-based representation, thereby enabling a direct comparison between regular grid-based and relational modelling approaches.

We focus our comparison on the Recall, as it measures the fraction of true signal events correctly identified by the model. In our case, this metric is particularly relevant since the expected signal yield is extremely small. Although optimising for Recall may lead to sub-optimal background rejection, it maximises the signal acceptance, which is crucial for ensuring sensitivity to rare physics signatures, where missing signal events would significantly reduce discovery potential.

As shown in table 1, the MoE model achieves the highest overall performance, with a Recall of 0.852. The GT model follows closely, reaching a Recall of 0.827 and outperforming the CNN baseline (0.824), while the GAT and GCN architectures perform less effectively. The superior performance of

the MoE highlights the advantage of combining multiple specialised experts and exploiting the full relational information of the calorimeter data, which is consistent with the need to model heterogeneous and irregular signal topologies. The comparatively lower performance of the GCN can be attributed to its fixed neighbourhood aggregation scheme, which does not dynamically weight neighbouring nodes according to their physical relevance; this is a limitation for heterogeneous calorimeter showers, where different energy deposits can contribute very differently to the classification task [22].

The strong results obtained with GT further confirm that graph-based representations can enhance sensitivity to rare signal topologies beyond what can be achieved with conventional image-based approaches, by preserving the underlying geometric and relational structure of the detector response.

6. Interpretability analysis

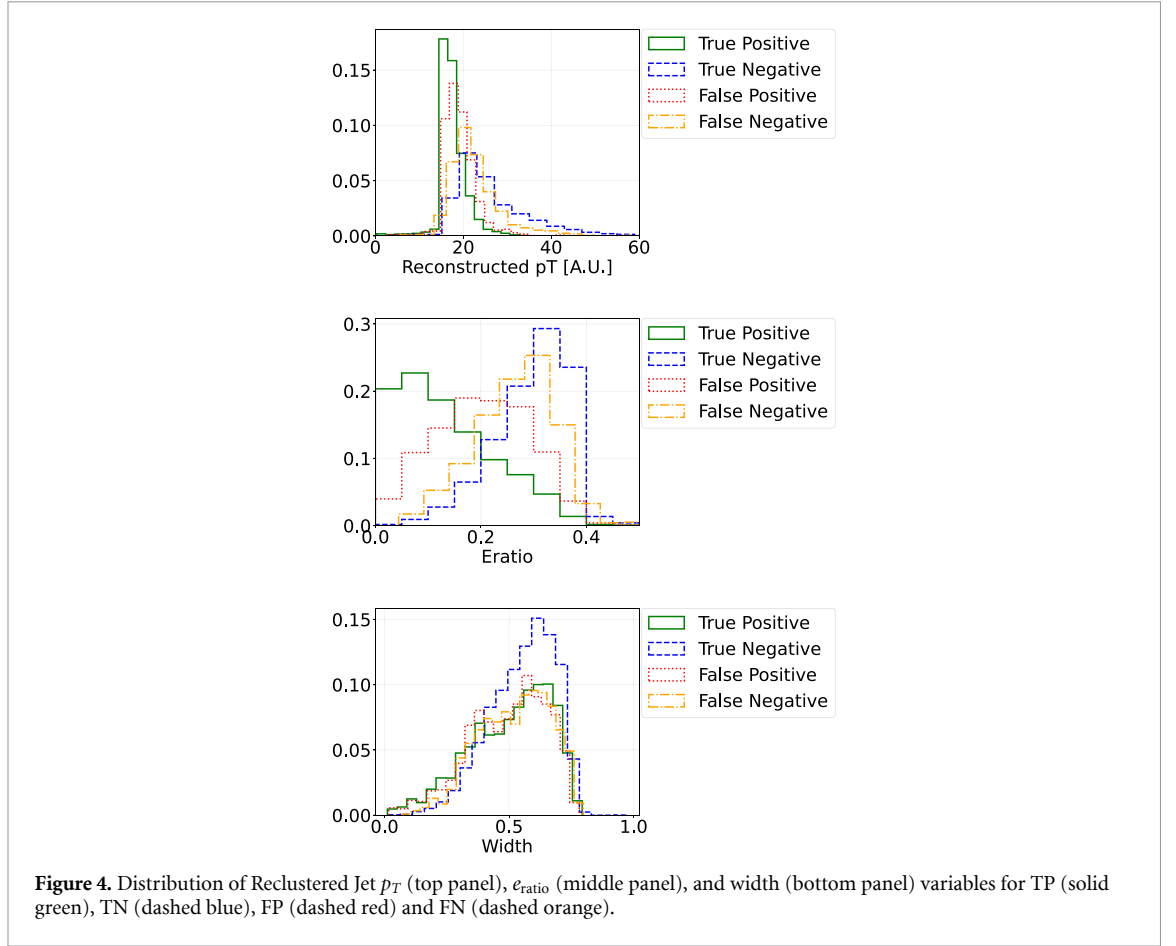
For the analysis of the results, the dataset is divided into four distinct categories based on the output predictions and their corresponding labels: true positives (TP), which are correctly identified signal graphs; true negatives (TN), which are correctly identified background graphs; false positives (FP), which are background graphs incorrectly classified as signal; and false negatives (FN), which are signal graphs incorrectly classified as background. The categorisation is performed by applying a threshold of 0.5 on the model output score, separating events predicted as signal ($\text{score} > 0.5$) from those predicted as background ($\text{score} \leq 0.5$) for the two classes, allowing a systematic comparison of model behaviour across correctly and incorrectly classified events. The threshold of 0.5 was selected as the standard symmetric decision boundary used for binary classification. Since optimising the threshold directly for Recall would lead to the trivial solution of classifying all events as signal, a threshold scan was instead performed by maximising the validation accuracy. This yielded an optimal threshold of 0.48 ± 0.03 , statistically consistent with the adopted value of 0.5.

Analysing physics quantities extrapolated from the graph can further help identifying interesting patterns, and provides a physics-grounded interpretation of the model predictions. The following quantities are reconstructed:

- reclustered p_T , a proxy for the jet momentum in the transverse plane;
- e_{ratio} , the ratio of the sum of the energy clusters in the first sampling layer to the total jet energy;
- width, the largest angular distance between two clusters on the same sampling layer.

The *reclustered* p_T for all categories is shown in figure 4(top panel). A clear pattern emerges when comparing correctly and incorrectly classified samples. For background events misidentified as signal (false positives), the reclustered p_T distribution tends to be shifted towards higher values relative to the true negatives, indicating that the model tends to associate higher-energy background jets with signal-like features. Conversely, signal events misclassified as background (false negatives) exhibit a lower p_T distribution compared to the true positives, suggesting that low-energy signal jets are more easily confused with background. This behaviour reflects the model's sensitivity to jet energy as a discriminating feature and highlights the overlap between low- p_T signal and high- p_T background events that contributes to the classification ambiguity, illustrating a physically meaningful source of misclassification. From a HEP perspective, this trend is consistent with expectations for long-lived particle decays, where the resulting displaced jets often carry lower transverse momentum and display broader, less collimated energy profiles, making them intrinsically more challenging to distinguish from standard background jets, thus indicating that the model captures relevant physical characteristics of the signal. For the jet width, shown in figure 4(bottom panel), the distributions are quite similar across all predicted categories. This is expected because jet width is a relatively simple variable for discrimination, and even a baseline classifier can separate signal from background effectively using it. From a physics point of view, the signal, which is originating from late decays, is expected to be narrower than the prompt background, since the electromagnetic or hadronic shower starts developing later in the calorimeter, producing a more collimated energy deposition. Thus jet width carries limited additional information for the classifier once other, more complex features are considered. This is consistent with the observation that misclassified jets (FP and FN) show only minor deviations in their width distributions compared to correctly classified jets, suggesting that this variable plays a secondary role in the model's decision process.

Distributions of the e_{ratio} variable are presented in figure 4(middle panel). The TP peak at low values of e_{ratio} , around ~ 0.1 , while TN and FN peak at higher values, around ~ 0.4 . False positives fall in between, around ~ 0.2 . This behaviour reflects the classifier's decision boundary and the underlying



physics: signal jets (TP) tend to have low e_{ratio} , consistent with late-depositing showers that concentrate energy in the last layers of the calorimeter. Background jets (TN) naturally peak at higher values, indicative of more diffuse, prompt energy deposition. False negatives inherit the background-like e_{ratio} distribution, explaining their peak near 0.4, while FP are intermediate because the classifier has overridden part of the background behaviour to mimic the signal, selecting events with moderately low e_{ratio} that partially resemble the signal pattern. This shows that e_{ratio} is a physically meaningful variable for distinguishing late-decay jets, and the network uses it in combination with other features to resolve ambiguous cases, highlighting its importance in the learned representation.

Figure 5 shows the correlation between the reclustered p_T and e_{ratio} for all prediction categories. A positive correlation is observed for the true signal events, with a correlation coefficient of about 0.21, which indicates that late-decay jets with higher reclustered p_T tend to exhibit slightly larger e_{ratio} values, although still concentrated at low values of e_{ratio} (around ~ 0.1), consistent with their compact and delayed energy deposition pattern. For the TN, FP, and FN, the correlation is weak. The TN and FN populations cluster around e_{ratio} 0.4, as expected for prompt energy deposits, while FP events occupy an intermediate region around e_{ratio} 0.2, reflecting the network's tendency to partially adapt background features toward the signal domain. The pattern confirms that e_{ratio} retains strong discriminating power at low values, and that the model primarily exploits this region in combination with reclustered p_T to identify signal-like behaviour, demonstrating a consistent interplay between energy scale and shower development in the learned features.

Finally, we also depict the correlation between the number of nodes and the reclustered p_T in figure 6. A positive correlation is observed for TP events, indicating that jets with higher reclustered p_T contain a larger number of nodes. This is consistent with physics expectations, where as the visible energy increases additional energy clusters are formed. A similar trend is seen for FP, reflecting that some background jets with higher p_T can mimic the signal's node multiplicity and be misclassified. For the true and FN the correlations are smaller, suggesting a more stochastic shower distribution for the prompt jets. These patterns show that the network captures the expected physical dependence between jet energy and topological complexity, while the relative strength of the correlations across categories indicates that node multiplicity is one of the features contributing to the discrimination of signal-like

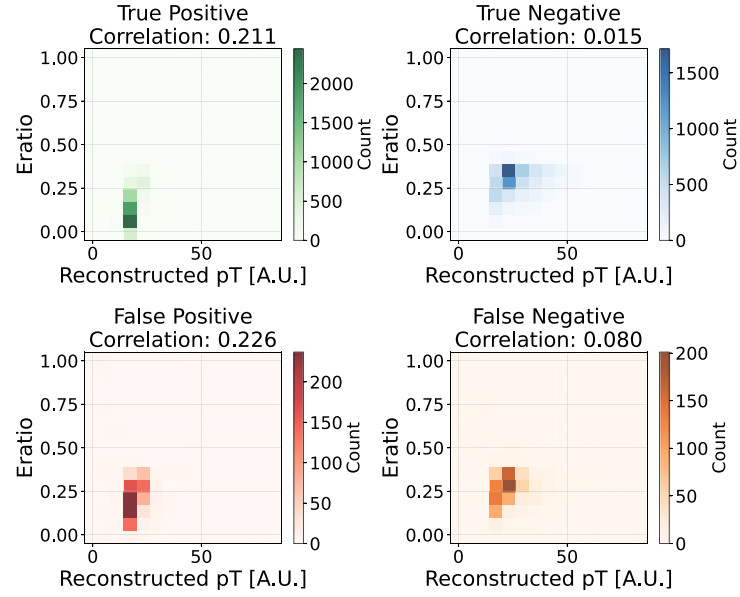


Figure 5. Correlation plots between p_T and e_{ratio} for MoE model.

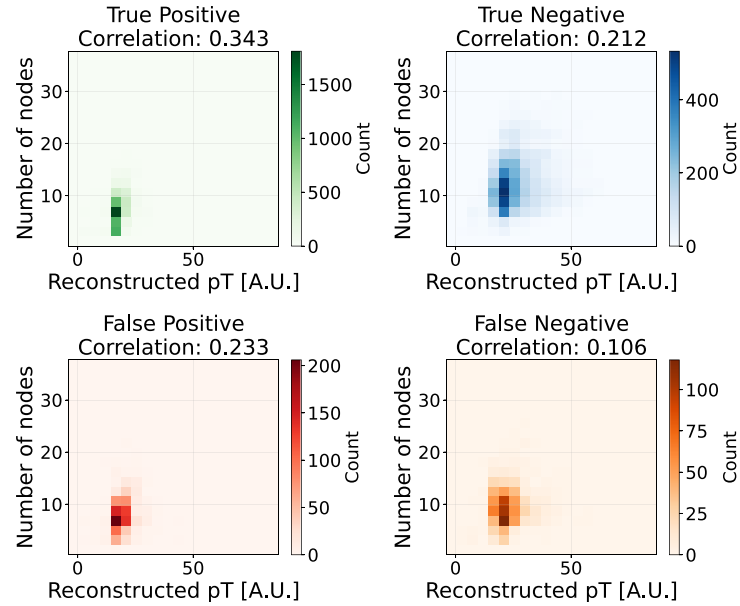


Figure 6. Correlation plots between p_T and the number of nodes for MoE model.

jets at high p_T , providing further evidence that the model encodes physically meaningful relationships between observables and classification outcomes.

Overall, the results demonstrate that the model captures physically meaningful relationships between calorimeter observables and classification outcomes. The observed dependencies between p_T , e_{ratio} , and node multiplicity are consistent with the expected behaviour of displaced and prompt jets, indicating that the model exploits relevant physical features rather than spurious correlations. This provides additional confidence in the robustness and interpretability of the proposed approach, particularly in the context of rare signal identification. Attention-based visualisations were explored but found to be difficult to interpret in this setting due to the nature of calorimeter nodes, which represent energy deposits rather than semantically distinct objects. A preliminary analysis of the expert routing (see figure 3 in appendix A) shows no strong differentiation between experts across event categories, suggesting that the performance gain is primarily driven by increased representational flexibility rather than strict expert specialisation.

7. Conclusion

This work investigated the behaviour and interpretability of a MoE model trained to discriminate BSM late-decaying jets from standard jets, with a focus on interpretability and understanding of the correlations among cluster level and jet-level. The distributions of key physics-motivated observables, such as the reclustered p_T , jet width, and e_{ratio} , are exploited to show how the network interprets physically meaningful patterns to separate signal from background. The study of individual features, as well as their correlations, provides insight into how the model refines its classification. Among the architectures studied, the MoE model consistently achieved the highest overall performance, outperforming both the baseline CNN and dense architectures. This improvement is consistent with the increased flexibility of the MoE framework, which enables a modular representation of heterogeneous jet topologies. The results indicate that the MoE classifier captures both local and global shower characteristics in a way that is consistent with HEP expectations, suggesting that its internal representations remain interpretable in terms of known detector and physics behaviour.

This analysis highlights the value of combining physics-driven observables with data-driven learning models. Understanding the interplay between cluster-level and jet-level features not only improves trust in the model's decision-making process but also provides guidance for the development of robust and interpretable approaches to rare signal identification in complex detector environments.

Acknowledgments

The work was partly funded by the European Union's CHIST-ERA programme under Grant Agreement CHIST-ERA-19-XAI-009 (MUCCA). SS is partly funded by Sapienza Grant RG123188B3EF6A80 (CENTS). DG is supported by PNRR MUR Project PE0000013-FAIR. The contribution of AD was carried out while enrolled as a PhD student at Sapienza University of Rome. AI-based tools were used for language editing and grammar improvement during the preparation of this manuscript.

Data availability statement

The data used is public and released under Creative Commons Zero v1.0 universal license. It is fully accessible via hepdata [16], where a description of the content is included. The generation and simulation setup are fully detailed in the cited ATLAS paper [13]. The code for the model is available at: <https://github.com/DonatellaGenovese/Transformer-for-DarkPhoton-Identification>.

All data that support the findings of this study are included within the article (and any supplementary files).

Supplementary data available at: <https://doi.org/10.1088/2632-2153/ae8e31/data1>.

ORCID iDs

Donatella Genovese  0009-0000-1865-1678

Alessio Devoto  0000-0002-7042-6061

Joseph Carmignani  0000-0002-1705-1061

Cristiano Sebastiani  0000-0003-1073-035X

Simone Scardapane  0000-0003-0881-8344

Monica D'Onofrio  0000-0003-2408-5099

References

- [1] Feickert M and Nachman B 2021 arXiv:2102.02770
- [2] ATLAS Collaboration (ATLAS) 2008 The ATLAS Experiment at the CERN Large Hadron Collider *J. Instrum.* **3** S08003
- [3] de Oliveira L, Kagan M, Mackey L, Nachman B and Schwartzman A 2016 Jet-images — deep learning edition *J. High Energy Phys. JHEP07(2016)069*
- [4] Madrazo C F, Cacha I H, Iglesias L L and de Lucas J M 2019 Application of a Convolutional Neural Network for image classification for the analysis of collisions in High Energy Physics *EPJ Web Conf.* **214** 06017
- [5] ATLAS Collaboration (ATLAS) 2017 Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1 *Eur. Phys. J. C* **77** 490
- [6] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C and Sun M 2020 Graph neural networks: a review of methods and applications *AI Open* **1** 57–81

- [7] Thais S, et al 2022 Graph neural networks in particle physics: implementations, innovations, and challenges (arXiv:2203.12852)
- [8] Qu H and Gouskos L 2020 Jet tagging via particle clouds *Phys. Rev. D* **101** 056019
- [9] ATLAS Collaboration 2026 Transforming jet flavour tagging at ATLAS *Nat. Commun.* **17** 541
- [10] Qu H, Li C and Qian S 2022 Particle transformer for jet tagging *Proc. 39th Int. Conf. on Machine Learning (Proc. Machine Learning Research)* vol 162, ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato (PMLR) (available at: <https://proceedings.mlr.press/v162/qu22b.html>) pp 18281–92
- [11] Birk J, Hallin A and Kasieczka G 2024 OmniJet- α : the first cross-task foundation model for particle physics *Mach. Learn.: Sci. Technol.* **5** 035031
- [12] Dwivedi V P and Bresson X 2021 A generalization of transformer networks to graphs AAAI'21 *Workshop on Deep Learning on Graphs: Methods and Applications* (arXiv:2012.09699)
- [13] ATLAS Collaboration 2024 Search for light long-lived neutral particles from Higgs boson decays via vector-boson-fusion production from pp collisions at 13 TeV with the ATLAS detector *Eur. Phys. J. C* **84**
- [14] Mu S and Lin S 2025 A comprehensive survey of mixture-of-experts: algorithms, theory and applications (arXiv:2503.07137)
- [15] Veličković P, Cucurull G, Casanova A, Romero A, Liò P and Bengio Y 2018 Graph attention networks *Int. Conf. on Learning Representations* (available at: <https://openreview.net/forum?id=rjXMpikCZ>)
- [16] ATLAS Collaboration 2024 Search for light long-lived neutral particles from Higgs boson decays via vector-boson-fusion production from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector *Eur. Phys. J. C* **84** 719
- [17] ATLAS Collaboration 2010 The ATLAS Simulation Infrastructure *Eur. Phys. J. C* **70** 823
- [18] Agostinelli S et al 2003 Geant4—a simulation toolkit *Nucl. Instrum. Meth. A* **506** 250
- [19] Genovese D, Sgroi A, Devoto A, Valentine S, Wood L, Sebastiani C, Scardapane S, D'Onofrio M and Giagu S 2025 Mixture-of-experts graph transformers for interpretable particle collision detection *Sci. Rep.* **15** 27906
- [20] Lim D, Robinson J D, Zhao L, Smidt T E, Sra S, Maron H and Jegelka S 2023 Sign and basis invariant networks for spectral graph representation learning *The 11th Int. Conf. on Learning Representations* (arXiv:2202.13013)
- [21] Kipf T and Welling M 2017 Semi-supervised classification with graph convolutional networks *Int. Conf. on Learning Representations* (available at: <https://openreview.net/forum?id=SJU4ayYgl>)
- [22] Shlomi J, Battaglia P and Vlimant J R 2020 Graph neural networks in particle physics *Mach. Learn.: Sci. Technol.* **2** 021001