ATLAS NAPOLI

ATLAS EXPERIMENT

INFN NAPOLI
Istituto Nazionale di Fisica Nucleare
Sezione di Napoli

Università degli Studi di Napoli Federico II

ANOMALY DETECTION SEARCH IN RUN-3 USING THE VV FULLY HADRONIC CHANNEL: STATUS UPDATE

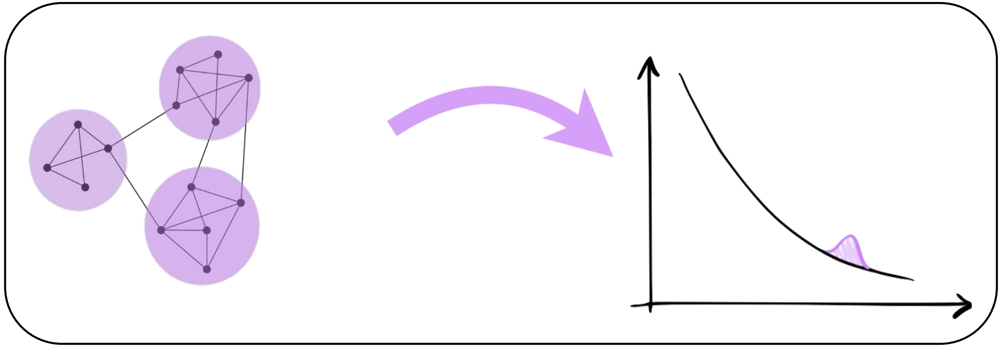
ANTONIO D'AVANZO, FRANCESCO CONVENTI, ELVIRA ROSSI, FRANCESCO CIROTTO

ATLAS group of Naples meeting, Napoli, 03/03/2023

IN THE PREVIOUS EPISODE

- The aim:
  - Graph Anomaly Detection algorithm for the discovery of diboson resonances decaying in fully hadronic final states with the ATLAS detector in run-III.
    - Anomaly Detection: **model-independent** approach, sensitive to more than one signal hypothesis as it only detects «anomalies» w.r.t. background.

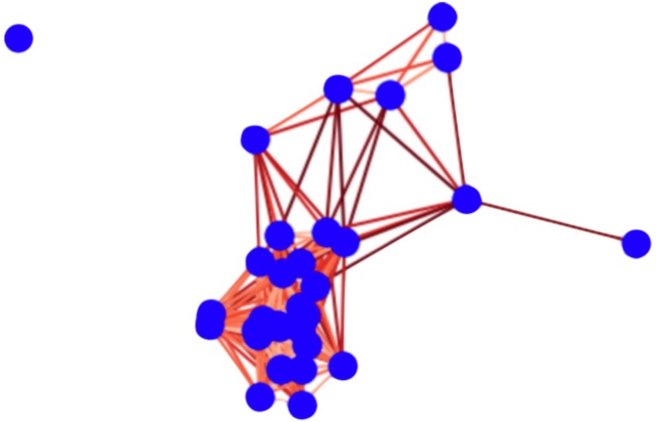
- The strategy:
  - Reconstructed jets at each event are represented as graphs.
  - This dataset of graphs is then used to train a Graph Neural Network (GNN) model in order to classify anomalous signal vs background by means of an anomaly score.



LAST STATE-OF-THE-ART

- Graph: Structured objects composed of entities used to describe and analyze relations and interactions (edges) between such entities (nodes).

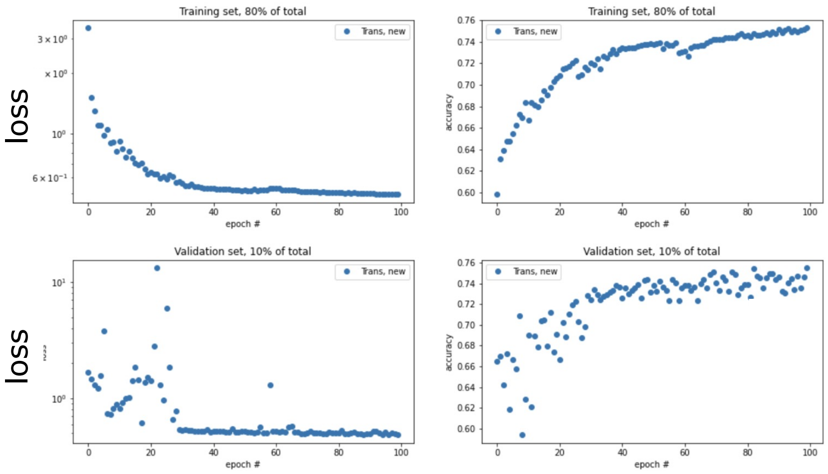
- **Definition of a jet**
  - Entites:
    - Nodes → topoclusters contained in each jet reconstructed with anti-k_t algorithm
    - Edges → Created only if $\Delta R < 0.4$ between two topoclusters, no self-loops
  - Features:
    - Nodes → pT fraction, η, ϕ.
    - Edges → $1/(\Delta R + \varepsilon)$

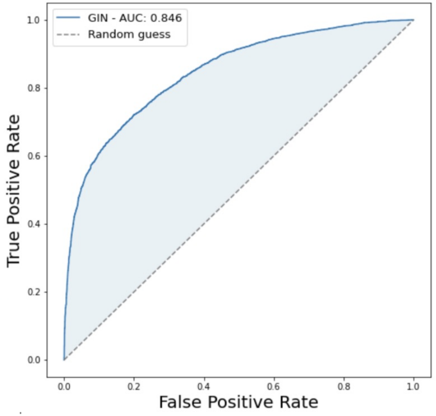


LAST STATE-OF-THE-ART: CLASSIFIER

- Graph Isomorphism Network trained on LHC Olympics 2020 R&D dataset for jet-level classification of anomalous sig vs bkg.
  - Supervised approach to provide baseline and check task doability.
  - Signal: Z' → XY → qqqq → JJ
  - Background: QCD di-jet.

Supervised training

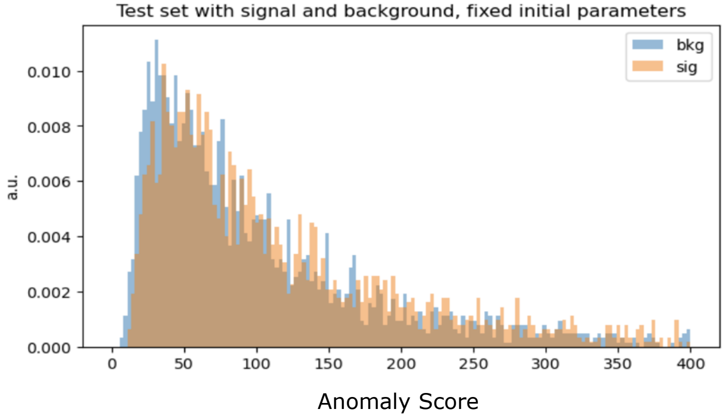




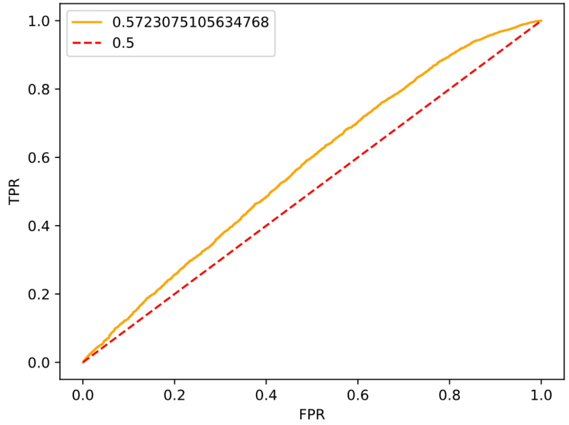
- Accuracy: **0.743**
- AUC: **0.832**

LAST STATE-OF-THE-ART: ANOMALY DETECTION

- Graph Isomorphism Network trained on LHC Olympics 2020 R&D dataset for jet-level classification of anomalous sig vs bkg.
  - Unsupervised way, actual anomaly detection algorithm.
  - Anomaly Score computed to decide wheter jet is signal or background.



Unsupervised training



- AUC: **0.572**

TEAM UPDATE

- Collaborative effort between the Napoli and Roma I ATLAS groups.
  - New addition to the team: master thesis student Antonio Corvino.
  - He will start working on the AD network in the next month.

Meet the full team

Who	Where	Role	FTE
Valerio Ippolito	Roma 1	Faculty	0.3
Stefano Giagu	Roma 1	Faculty	0.1
Graziella Russo	Roma 1	PhD (expected 2025)	1.0
Francesco Conventi	Napoli	Faculty	0.1
Elvira Rossi	Napoli	Faculty	0.1
Francesco Cirotto	Napoli	Post-doc	0.3
Antonio D'Avanzo	Napoli	Post-master fellow	1.0
Antonio Corvino	Napoli	Master student	1.0

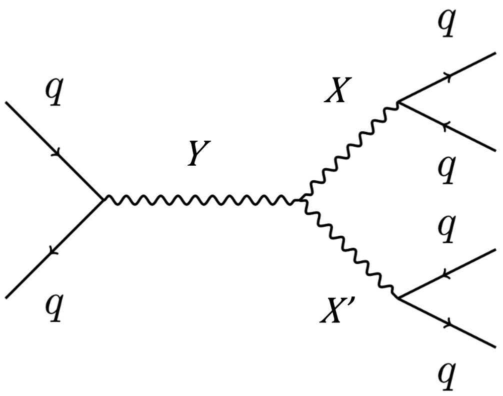
New!

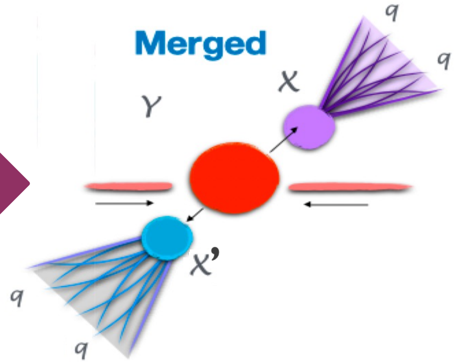
ANALYSIS CHANNEL OF INTEREST

- Analysis received kick-off in recent DBL meeting! Glance page also created.
  - Defined physics channel of research, to be conducted in the scope of DBL analyses.

- Events signature
  - Background: QCD multi-jet, to be estimated with data-driven approach (typical of unsupervised tasks).
  - Signal: Y → XX' → qqqq, where X or X' can be either SM or BSM bosons.

- Aiming at the kinematic region where X and X' are produced with significant Lorentz boost due to large mass difference between parent and daughters particles.
  - Y reconstructed with two large-R jets (fatjets) of radius ΔR, defined as **merged regime**.





Mass of decaying particle

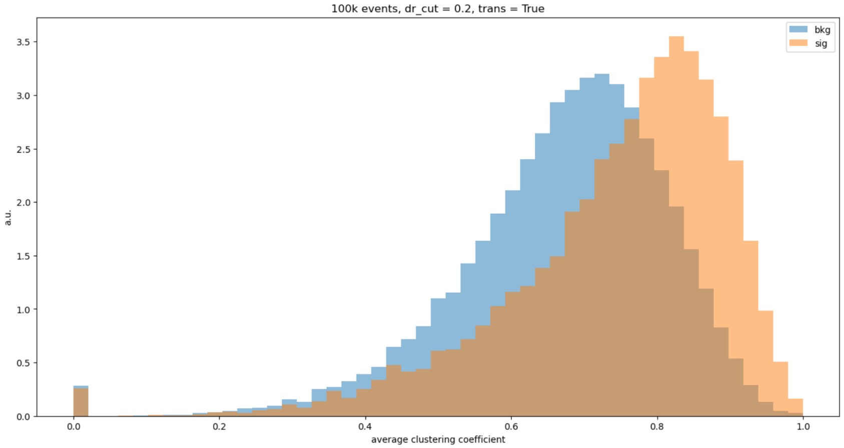
$$\boxed{\Delta R \cong \frac{2m}{p_T}}$$

p_T of decaying particle

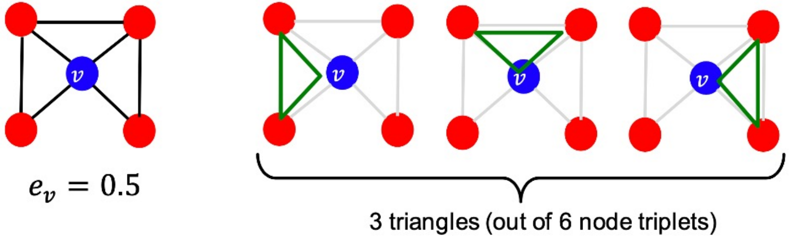
GRAPH REPRESENTATION UPDATE

- Change to the cut on dr between two topoclusters to decide wheter they are connected or not.
  - $\Delta R < 0.4 \longrightarrow \Delta R < 0.2$, doing so enhances the structural diversities of signal and background.

- On-going studies to find new node features to add as input to the network ([pTfrac, η, ϕ] at the moment).
  - Main focus on the average clustering coefficient (e_v) computed for each graph, D2 and τ jet variables next.
  - Need to understand how to incorporate global quantities as node features for the GNN.



$$e_v = \frac{\#(\text{edges among neighboring nodes})}{\boxed{\binom{k_v}{2}}} \in [0,1]$$



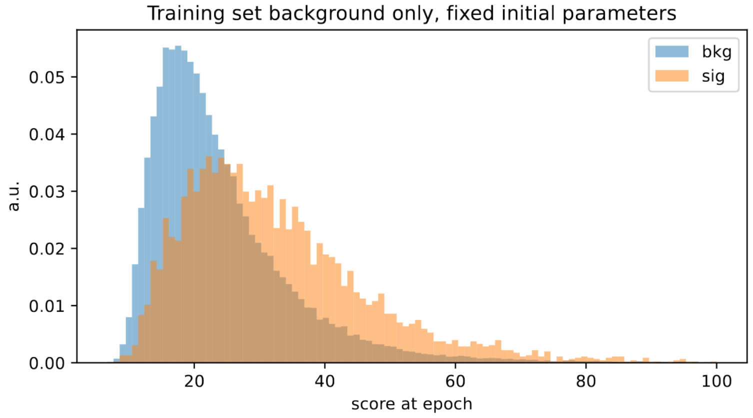
Jure Leskovec et al., CS224W: Machine Learning with Graphs Stanford Course, Fall 2021

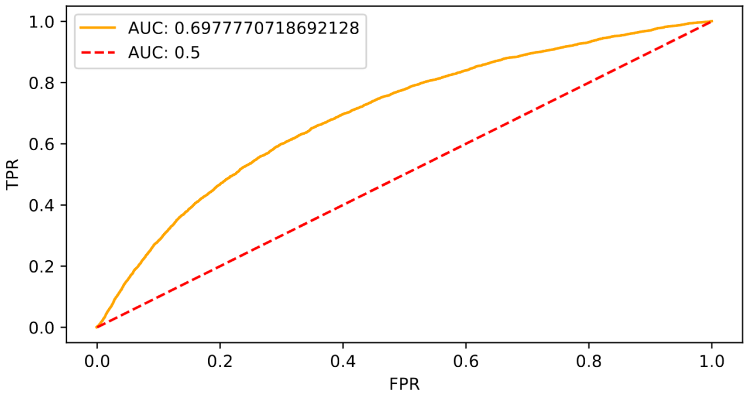
ANOMALY DETECTION GNN ARCHITECTURE UPDATE

- Graph vectorial representation $\hat{y}_G$ is obtained by pooling procedure from output of GNN, which is the vectorial representation h_v for each node of a graph.

$$\hat{\mathbf{y}}_G = \underbrace{\mathrm{Sum}}_{\text{Pooling function}}(\{\mathbf{h}_v^{(L)} \in \mathbb{R}^d, \forall v \in G\}) \xrightarrow{\text{new!}} \hat{\mathbf{y}}_G = \underbrace{\mathrm{Max}}_{\text{Pooling function}}(\{\mathbf{h}_v^{(L)} \in \mathbb{R}^d, \forall v \in G\})$$

- Consequent increase in performance, also aided by tuning the number of GIN layers and change of ΔR cut.



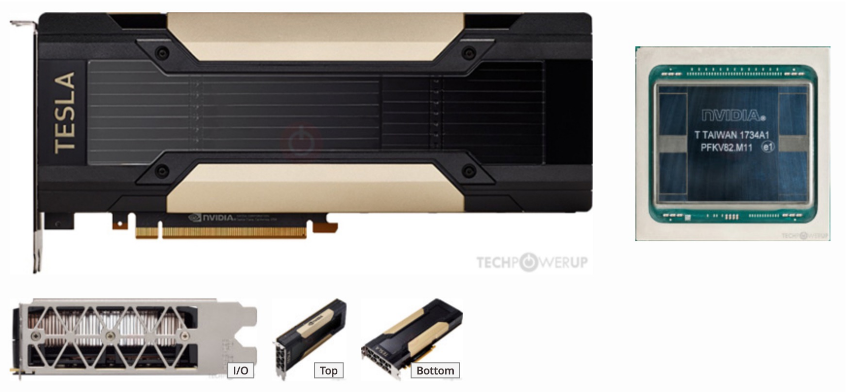


COMPUTATIONAL ENVIRONMENT UPDATE

- GNN development done on hardware provided with 2 GPUs (INFN I.Bi.S.Co cluster), although last time we were running trainings on CPU.
  - Now parallel training on one GPU, reducing computation time by 1/2 or 1/8 based on the used algorithm.
  - Could still gain from the second GPU, but complex changing in code is needed.

- Currently working on only one node of the cluster
  - Parallelization to all nodes at dispose is possible, work in progress with the help of Bernardino Spisso.

- 2x NVIDIA Tesla V100S PCIe 32 GB, 5120 cores
- Cluster composed of 6 nodes.



CONCLUSION

- Many changings applied w.r.t. last group I meeting.
  - New addition to the team from Naples side (Antonio Corvino).
  - Defined physics channel of interest ($Y \rightarrow XX' \rightarrow JJ$).
  - Changing of the value of graph connection criteria, on-going studies on new node features.
  - Increase in AD algorithm performance by ~12% thanks to max pooling function and fine tuning.
  - Computational time reduced by running GNN training on GPU.

- Next steps:
  - About the model:
    1. Testing GNN performance with other supervised (random forest, transformer) and unsupervised (autoencoder) techniques (WIP from Graziella Russo).
    2. Testing new Loss functions for maximising network performances (OCGTL, GLocalKD).
    3. Use of multiples nodes of I.Bi.S.Co cluster to run training, also both GPUs per nodes.
  - About the dataset:
    1. Inclusion of full detector info and event-based anomaly score, studies on new input features.
    2. Migration to real dataset to explore run-III data gathered by the ATLAS detector for new DBL searches or rediscovery of known resonances (WIP from Francesco Cirotto).

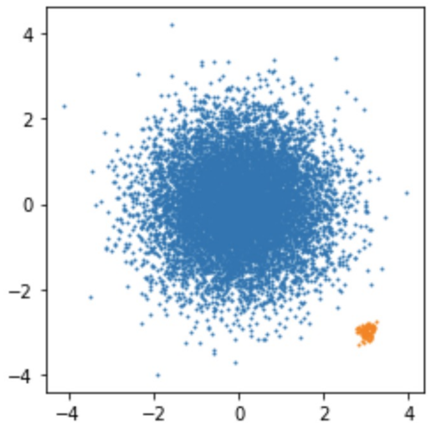
Thank you for your attention!

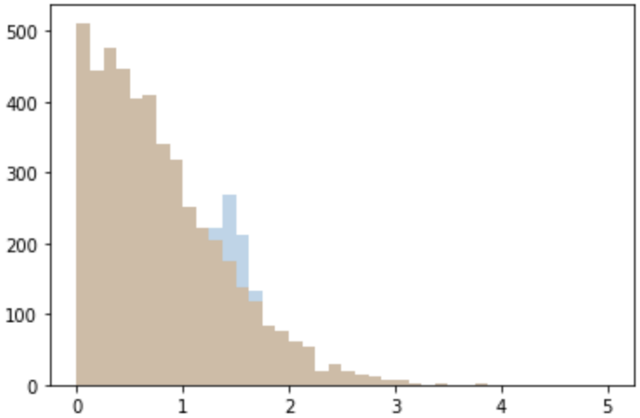
BACKUP

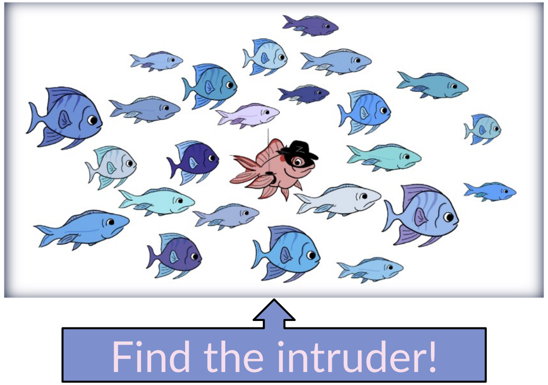
WHAT IS ANOMALY DETECTION?

- Anomalies: abnormal objects significantly different from other members of a sample.
  - Anomaly Detection refers to ML techniques used to spot these outliers.

- Particle physics scenario → Identification of **model-independent** features of detector data inconsistent with the expected background.
  - Related works: PRL 125 131801, arXiv:2105.09274, ATLAS-CONF-2022-045.

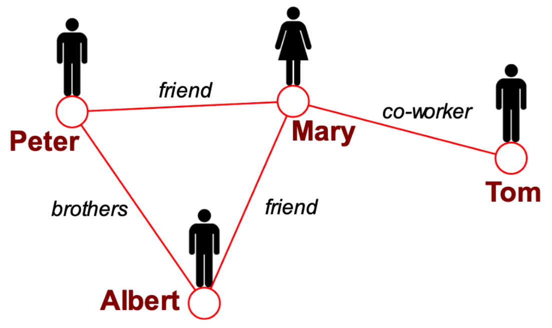




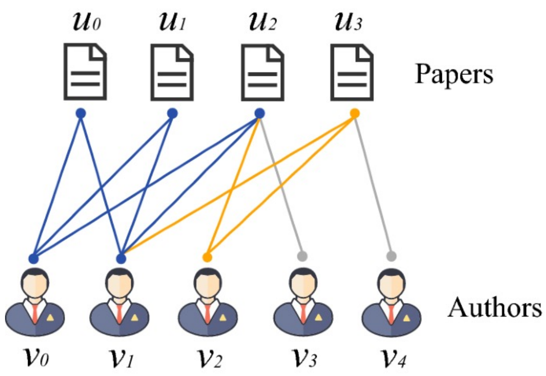


WHAT ARE GRAPHS?

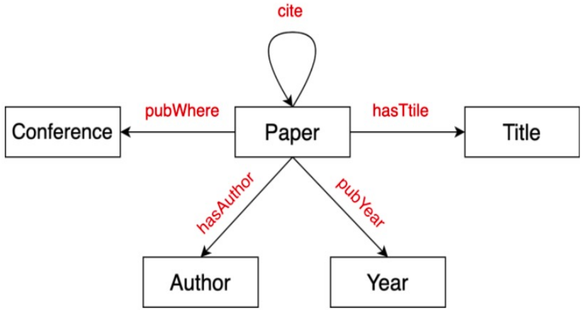
- Structured objects composed of entities used to describe and analyze relations and interactions (edges) between such entities (nodes).
  - Nodes and edges typically contain features specific to each element and each pair.
  - Many types of graphs based on the relations: directed, heterogeneous, bipartite, weighted ecc.



Undirected graph



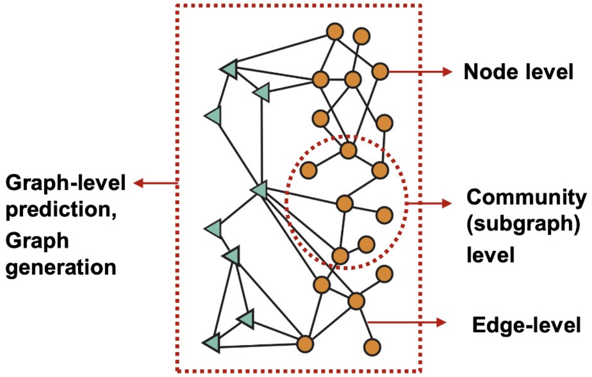
Bipartite graph



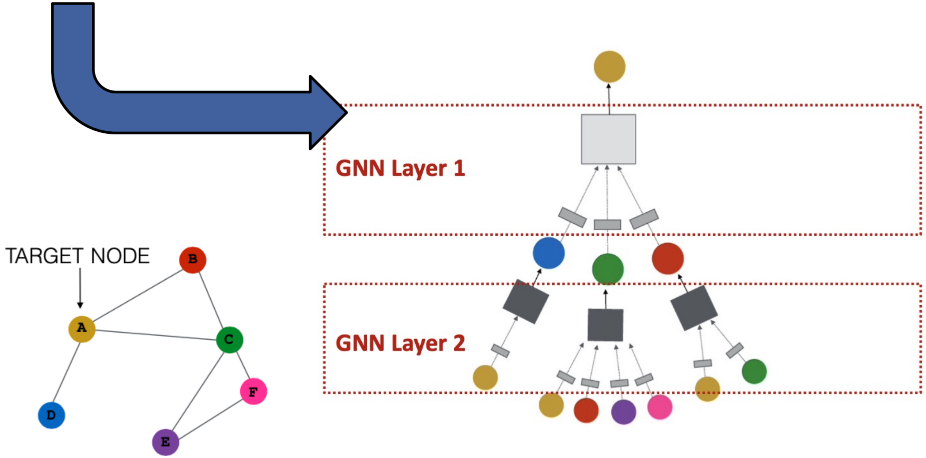
Directed graph

GRAPH NEURAL NETWORKS

- Graph Neural Networks (GNNs) are ML architectures built specifically to make predictions on graphs, exploiting their relational nature.
  - Based on learnt vector representation (embedding) of each node of the input graphs.
- The embeddings are updated at each layer by aggregating the information passed between the target node and the nodes from its closest neighbourhood → **message passing**



Several task levels, carried out by processing the final node embeddings in certain ways.



INPUT GRAPH

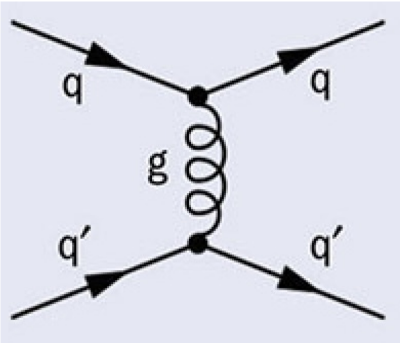
Each layer of GNN extends the neighbour range

CURRENT DATASET

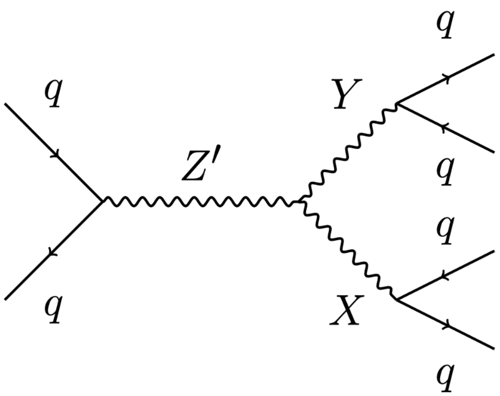
- Benchmark application with LHC Olympics 2020 R&D dataset.
  - MC generated dataset built specifically for anomaly detection.
  - 1.1M total events, 1M background and 100k anomalous signal.

- Events signature
  - Background: QCD di-jet.
  - Signal: Z' → XY → qqqq, particles reconstructed as fatjets with large radius R = 1.

QCD di-jet

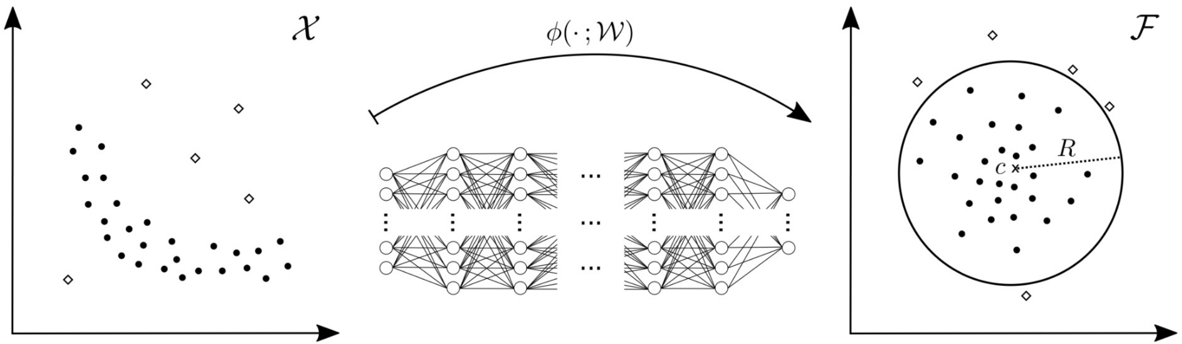


Particle	Mass [GeV]
Z'	3500
X	500
Y	100



DEEP SUPPORT VECTOR DATA DESCRIPTION (DEEP SVDD)

- Deep SVDD works by minimizing an objective in order to learn and optimize the radius R of a hypersphere in the output space F which only cointains outputs from non-anomalous data features X.
  - Output space defined by the output of the considered ML architecture (NN, MLP, GNN, ecc.)
  - Output from anomalies falls outside of the hypersphere and is identified by its distance from the center c.



objective

$$\min_{W} \frac{1}{N}\sum_{i=1}^{N} \|\mathrm{GIN}(G_i;W) - \mathrm{c}\|^2 + \frac{\lambda}{2}\sum_{l=1}^{L}\|W^l\|_F^2 \qquad W = \{W^1, \dots, W^l\}$$

$$s(\boldsymbol{x}) = \|\phi(\boldsymbol{x};\mathcal{W}^*) - \mathbf{c}\|^2$$

Anomaly Score

GRAPH ISOMORPHISM NETWORK (GIN)

- GIN formulation employs both message passing and MLPs, making it the most expressive GNN:

$$\mathrm{MLP}_{\Phi}\left((1+\underset{\substack{\downarrow\\ \text{learnable parameter}}}{\epsilon}) \cdot \mathrm{MLP}_f(c^{(k)}(v))) + \sum_{u \in N(v)} \mathrm{MLP}_f(c^{(k)}(u))\right)$$

$$\boxed{c^{(k)}(u) \leftrightarrow h_j^{(l)}}$$

Embedding of node u al layer (k)

- This expression can be rewritten in a more general way, also allowing for edge weights to be considered in the graph convolution.

$$\boxed{h_i^{(l+1)} = f_{\Theta}\left((1+\epsilon)h_i^l + \text{aggregate}\left(\left\{e_{ji}h_j^l, j \in \mathcal{N}(i)\right\}\right)\right)}$$

- Aggregate can be any permutation invariant function (Sum, Mean, Max ecc.)

GLOCALKD

- Article: Deep Graph-level Anomaly Detection by Glocal Knowledge Distillation.

- Employs a variation of Knowledge Distillation (KD) technique, where the initial goal is to train a simple model that syntetize the knowledge of a large model while maintaining similar accuracy as the large model.
  - Random Knowledge Distillation for joint distillation at node-level and graph-level.

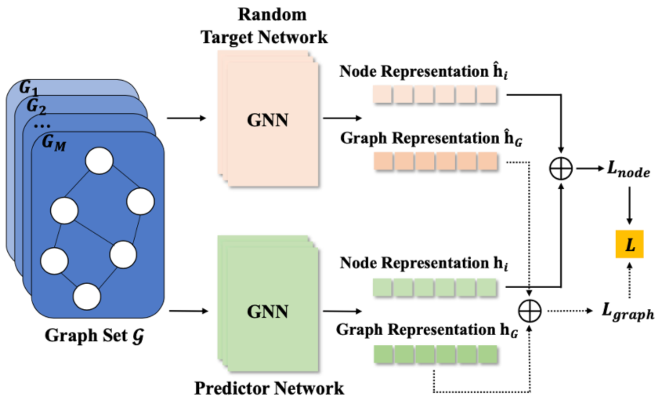
- Implements two GNNs:
  - Random Target Network, not-trained and randomly initialized, used as reference to learn the normal patters of our dataset.
  - Predictor Network, trained by comparing its node and graph representations $(\mathrm{h_G}, \mathrm{h_i})$ with the ones from the above network $(\widehat{\mathrm{h_G}}, \widehat{\mathrm{h_i}})$ through a KD function.

Loss function for unsupervised learning:

$$L = L_{graph} + \lambda L_{node}$$

$$L_{graph} = \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} KD\left(\mathbf{h}_G, \hat{\mathbf{h}}_G\right),$$

$$L_{node} = \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \left(\frac{1}{|G|} \sum_{v_i \in \mathcal{V}_G} KD\left(\mathbf{h}_i, \hat{\mathbf{h}}_i\right) \right)$$



GLOCALKD PAPER MODEL

- KD function in L chosen as error between the two networks output.

$$L = L_{graph} + L_{node}$$

$\lambda = 1$

$$L_{graph} = \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \|\mathbf{h}_G - \hat{\mathbf{h}}_G\|^2,$$

$$L_{node} = \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \left(\frac{1}{|G|} \sum_{v_i \in \mathcal{V}_G} \|\mathbf{h}_i - \hat{\mathbf{h}}_i\|^2 \right)$$

- Anomaly Score computed for test dataset:

$$f(G; \hat{\Theta}, \Theta^*) = \left\|\mathbf{h}_G - \hat{\mathbf{h}}_G\right\|^2 + \frac{1}{|G|} \sum_{v_i \in \mathcal{V}_G} \|\mathbf{h}_i - \hat{\mathbf{h}}_i\|^2$$

G → number of nodes for graph $\mathcal{G}$

- Things to note:
  - Node degree used as node feature for training.
  - Max pooling to obtain graph representation.

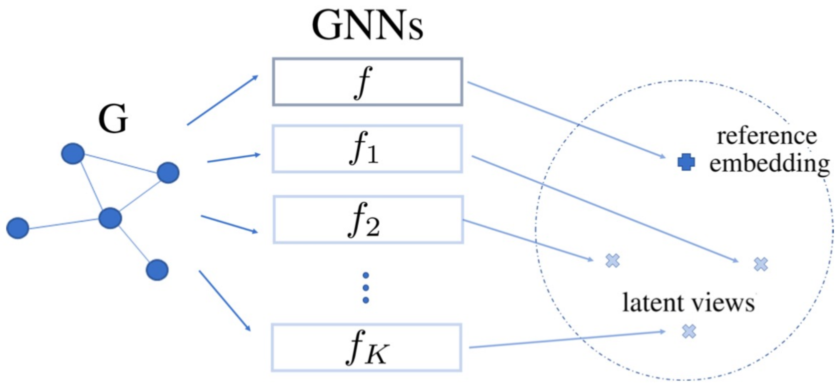
OCGTL

- Article: Raising the Bar in Graph-level Anomaly Detection.

- Combines one-class classification of OCGIN and neural transformation lerning.

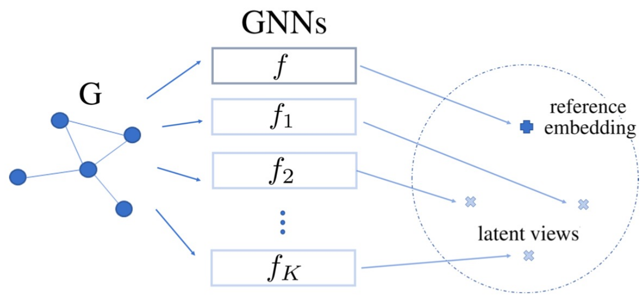
- One reference GNN and K additional GNNs are trained together in order to detect anomalies.
  - Each representation obtained now is used to learn the optimal hypersphere radius of non-anomalies region.

- Advantage w.r.t. DeepSVDD objective:
  - No hypersphere collapse, center can be treated as learnable parameter.
  - More robust training.
  - No performance flip.



OCGTL LOSS FUNCTION

- Consists of two terms:



$$\mathcal{L}_{\text{OCGTL}}(\text{G}) = \mathcal{L}_{\text{OCC}}(\text{G}) + \mathcal{L}_{\text{GTL}}(\text{G})$$

$$\boxed{\mathcal{L}_{\text{OCC}}(\text{G}) = \sum_{k=1}^{K} \|(f_k(\text{G}) - \theta)\|_2}$$

One class contribute

$$\mathcal{L}_{\text{GTL}}(\text{G}) = -\sum_{k=1}^{K} \log \frac{c_k}{C_k}$$

$$c_k = \exp\left(\frac{1}{\tau}\text{sim}(f_k(\text{G}), f(\text{G}))\right),$$

$$C_k = c_k + \sum_{l \neq k}^{K} \exp\left(\frac{1}{\tau}\text{sim}(f_k(\text{G}), f_l(\text{G}))\right)$$

Transformation learning contribute

- $\theta \rightarrow$ center of the hypersphere, sim chosen as cosine similarity $\rightarrow z^T z' / \|z\| \|z'\|$
- τ temperature parameter, final loss on training dataset at each epoch computed as $\mathbb{E}_{\text{G}}\left[\mathcal{L}_{\text{OCGTL}}(\text{G})\right]$