Anomaly detection search in fully hadronic final state

Francesco Cirotto
Università degli Studi di Napoli Federico II

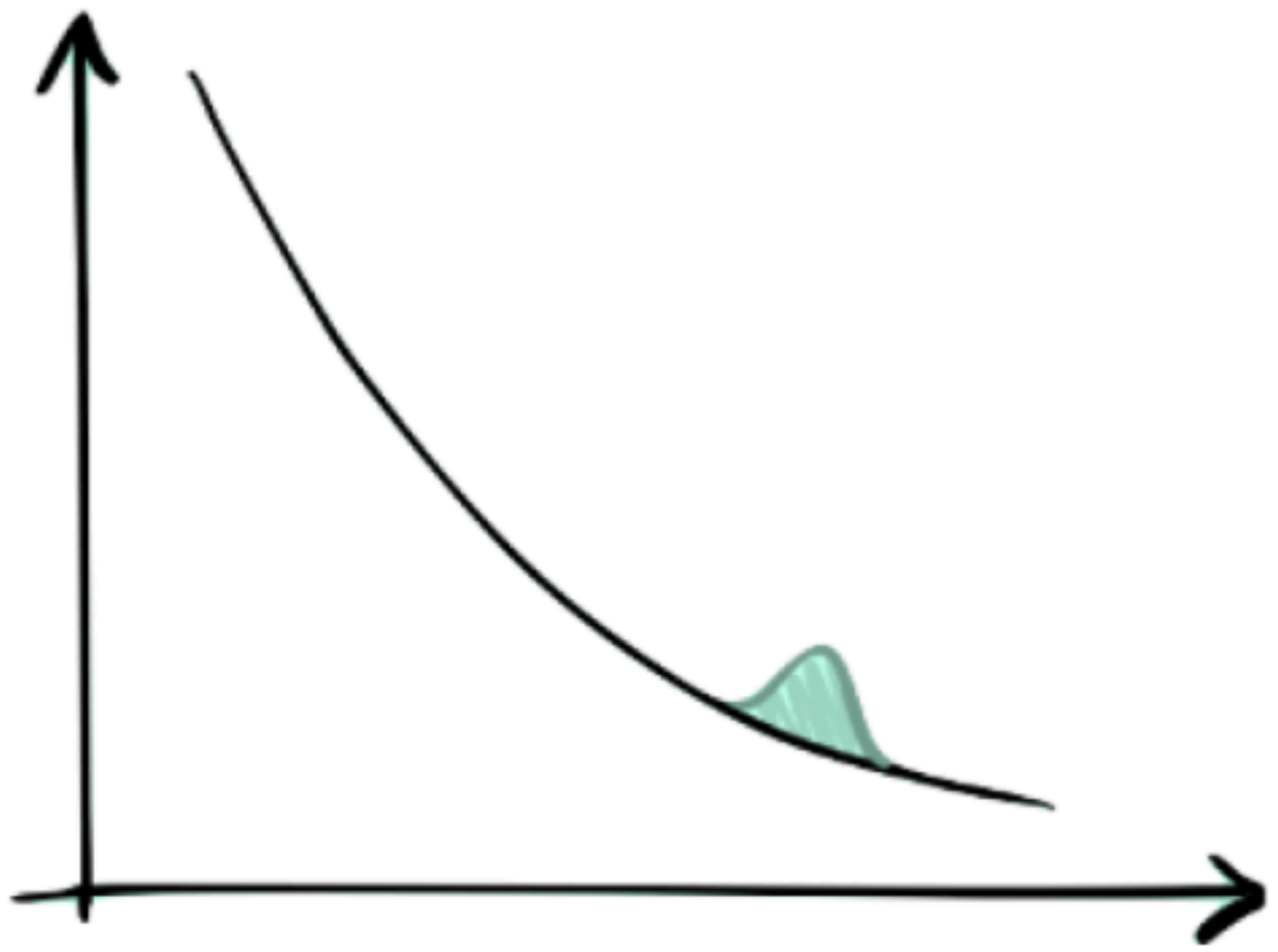
WHERE IS LHC GOING?

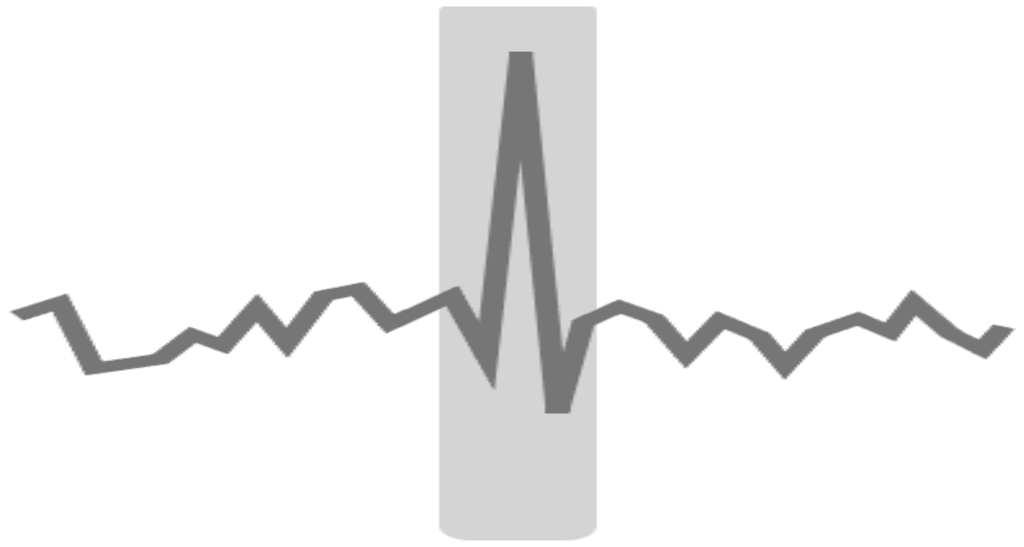
Finding anomalies in data

- No Physics Beyond Standard Model (BSM) has been observed at the LHC (yet!)

- The currently most used search paradigm is using model-dependent approaches
  - ↪What if these models have blind spots for unconventional new physics signatures?
  - ↪If there's new Physics in the current LHC data we can't miss it!

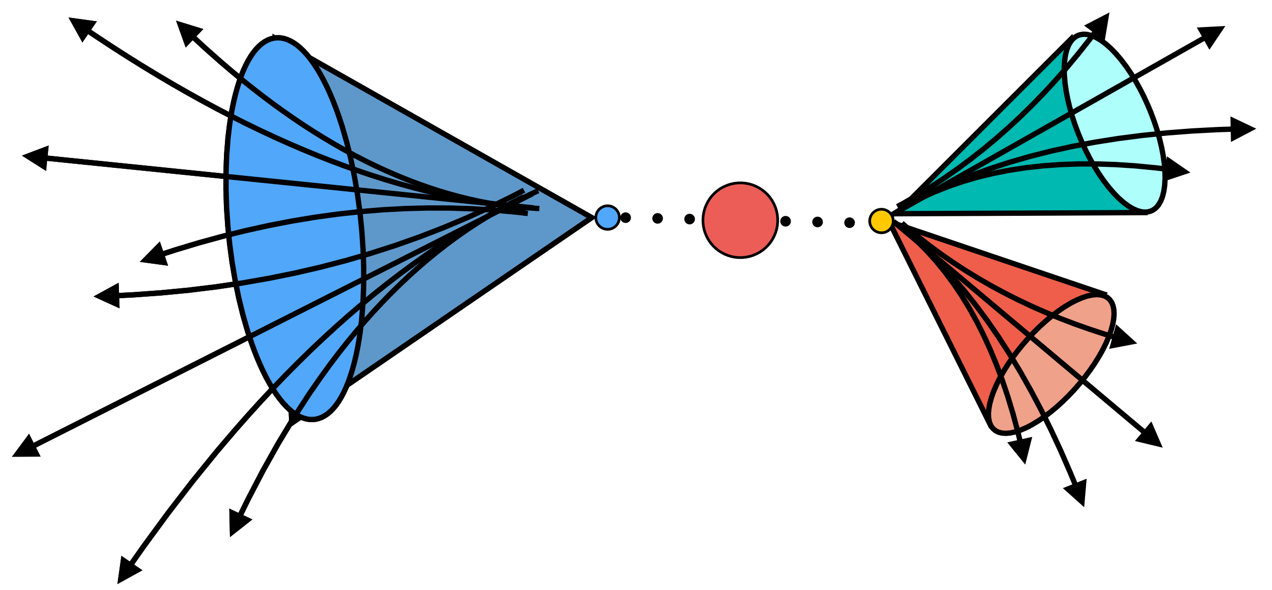
- Anomaly Detection (AD) uses unsupervised Machine Learning architectures to identify outliers in a set of "standard" objects.
  - ↪In High Energy Physics, this means the identification of features of detector data inconsistent with the expected background.



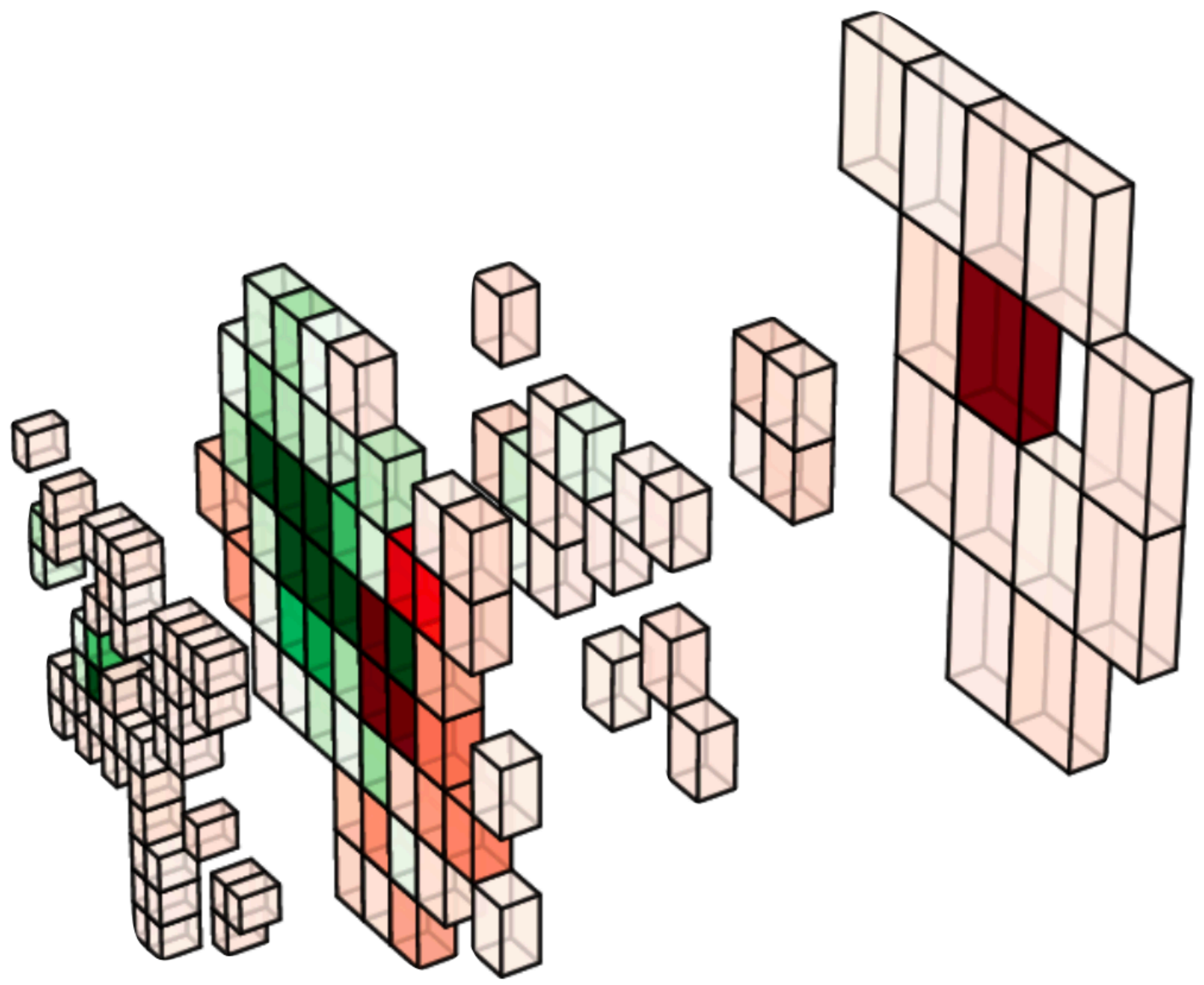


USING JETS IN AD

Jets as tools!



- Many Beyond Standard Model theories predict new massive resonances which can decay hadronically, leading to final states involving jets.
- For massive particles, their decay products become collimated, or 'boosted', in the direction of the progenitor particle.
  - ↪It is advantageous to reconstruct their hadronic decay products as a single large-radius (large-R) jet.



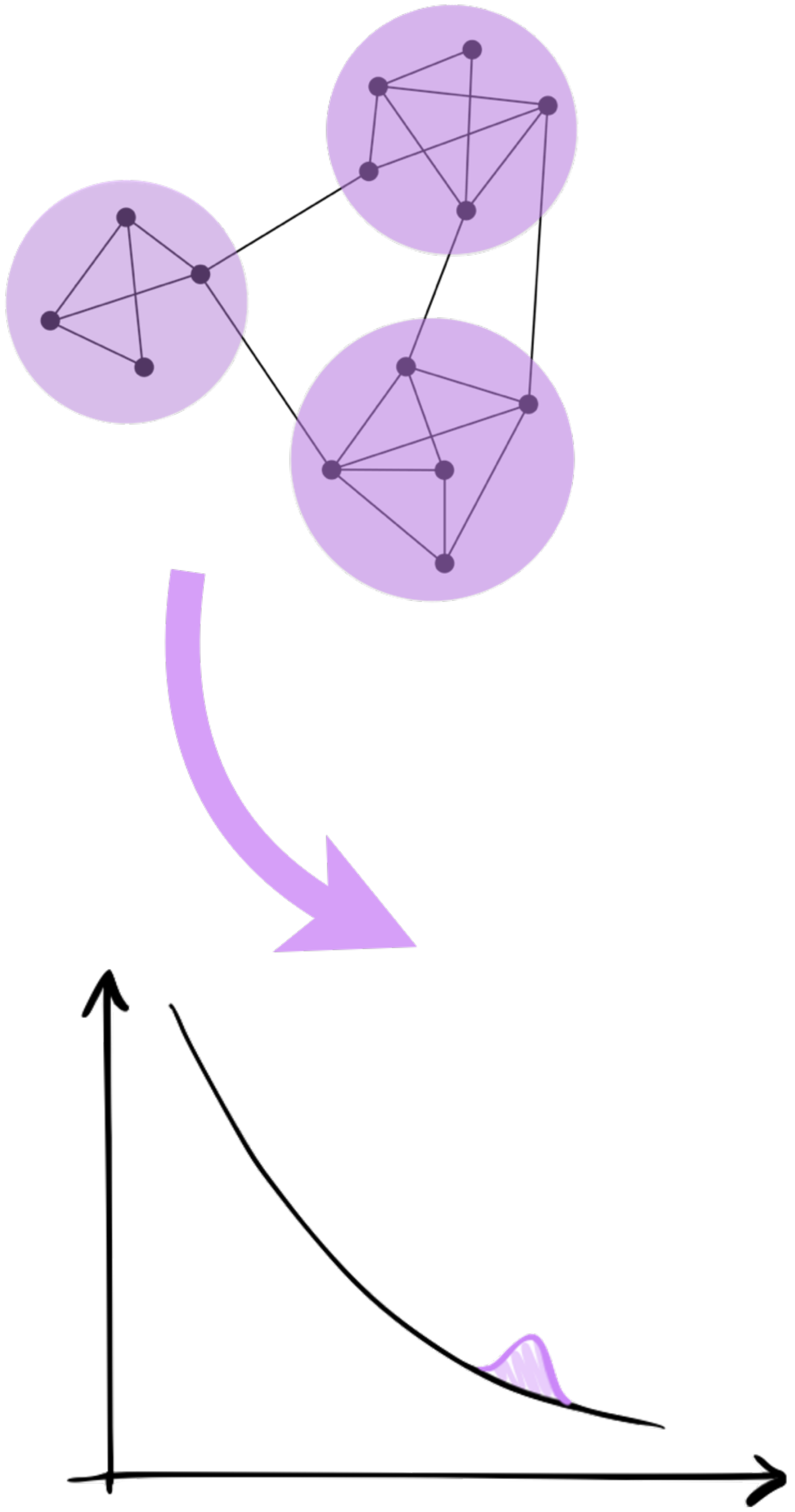
- Jet information can be used as input features for neural network architectures.
  - ↪A significant improvement in performances can be achieved by employing a set of features with basic information (low-level) such as information coming directly from the detectors.
  - ↪Jet constituents represent challenging input features to achieve this goal

The Idea

Graph Anomaly Detection for New Physics Searches

- Graph-structured data are ubiquitous across science, engineering, and many other domains
  - ↪Used to describe and analyze relations and interactions
  - ↪Can encapsulate object or event information
  - ↪Can be employed in particle physics!

- Our strategy: to represent jets as graphs and then apply machine learning to build an anomaly detection algorithm
  - ↪Targeting heavy resonance searches with hadronic final states in Run-3
  - ↪Exploit event-based graphs to detect anomalies



THE ANALYSIS IN ATLAS

- Find anomalies in our data with GNN
  - ↪Build graphs from jets
- From a jet level (already existing) to an event level approach
  - ↪Deviation from known SM processes
- Testing our model: apply the technique to other benchmark models
  - ↪Rediscovering "old" resonances as new anomalies! (W/Z/top?)

JETS AD GRAPHS

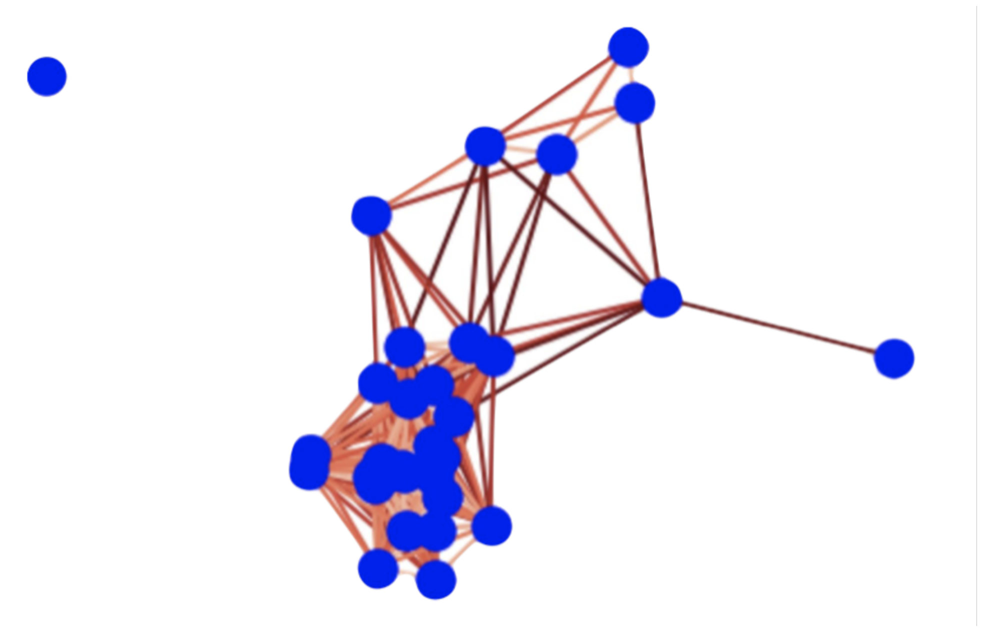
Graph definition

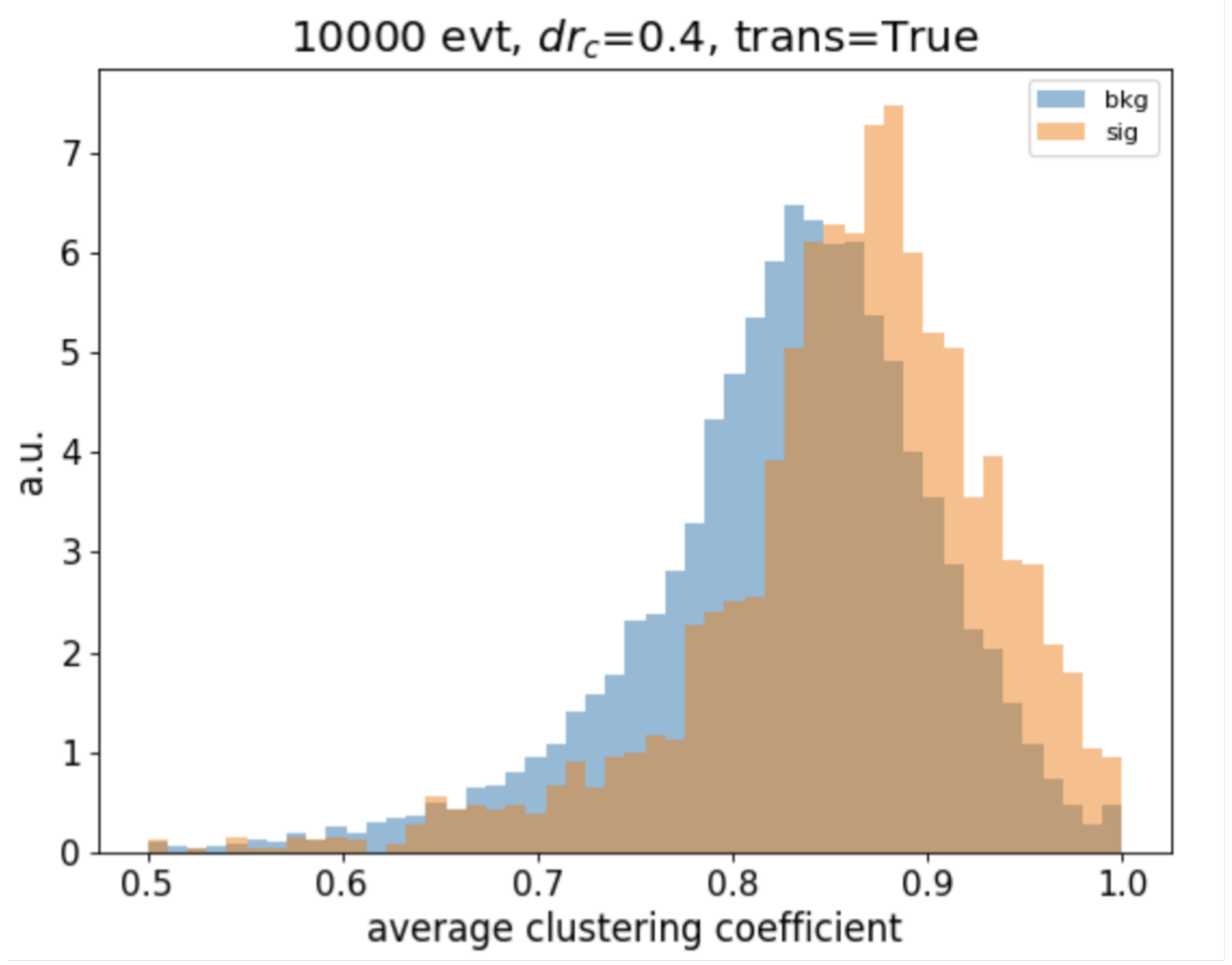
Features

- What is a node?
  - ↪Using jet constituents
  - ↪Fraction of jet pt, eta, phi

- What is an edge?
  - ↪Weight message from neighboring nodes
  - ↪Using "distance" between jets

- How are they connected?
  - ↪No self loops, DR cut = 0.4



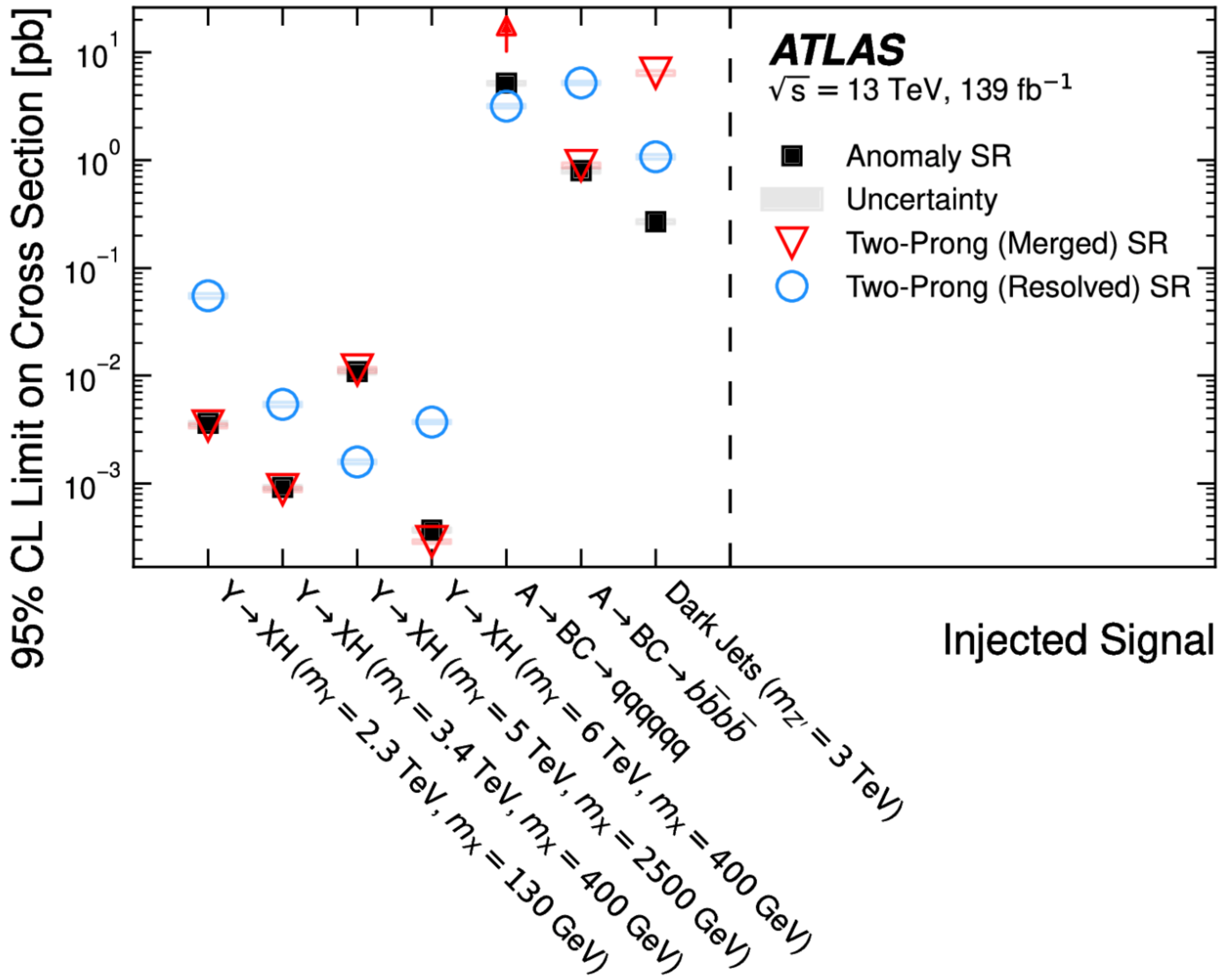


Measure of the degree to which nodes tend to cluster together

ATLAS Analysis

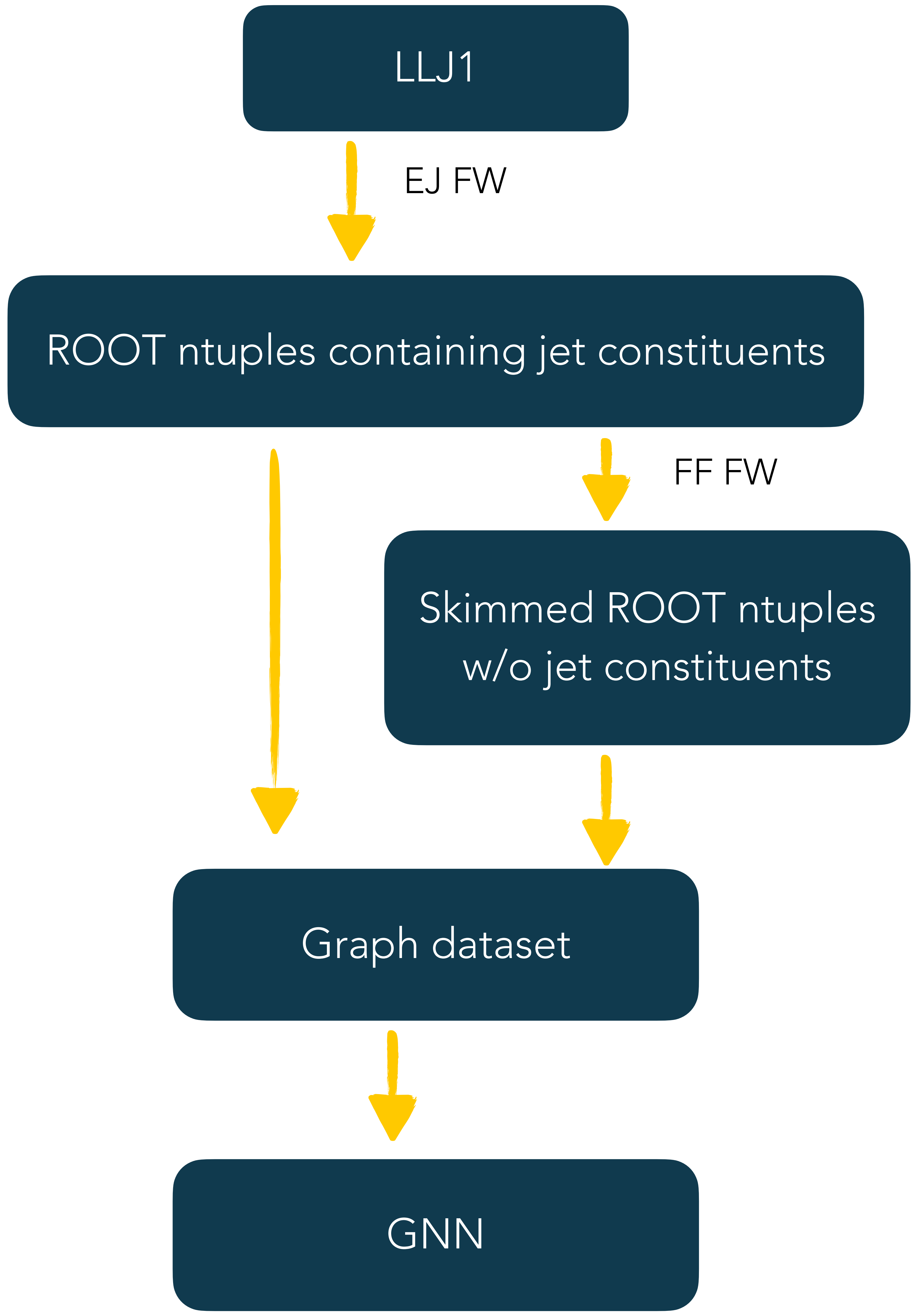
DATASETS

- Requested dedicated derivations (LLJ1) containing jet constituent information
  - ↪Due to the huge size agreed to not have MC simulations for all data taking periods
- Data 2022 and 2023
- MonteCarlo: mc23d campaign, which simulates 2023 data taking
  - ↪QCD main background divided in pt slices
  - ↪Top, V+jets
- Signals: several benchmarks signals.
  - ↪Heavy Vector Triplet: VVJJ, YXH
  - ↪Dark Jet
  - ↪3-prong signals
  - ↪Use only benchmark models, we do not want to cover the whole phase space



DATA PROCESSING WORKFLOW

- Two main steps in dataset processing:
  - ↪From LLJ1 derivations to ROOT ntuples via **EasyJet FW**
    - Can be already used for preliminary plots
    - A first preselection on DxAOD can be applied
  - ↪Skim EasyJet ROOT ntuples via **FastFrames FW**
    - Lighter ntuples
    - Dealing with events weights
    - Compute more complex variable/selection
- Create graph dataset for our ML architectures
  - ↪Use jet constituents to build graph



DATA PROCESSING WORKFLOW

Processing time

- EasyJet FW: run on grid, ~1 week
  - ↪Complete production is ~2 Tb
- FastFrames: run locally or via Condor ~0.5/1 day
  - ↪To be tested on Condor
  - ↪Complete production without constituents: ~33Gb
  - ↪Can be useful include jet constituents?
- Graph dataset creation: main bottleneck is computing time and dataset size
  - ↪~500 ev/s, ~2M ev/h depending on information stored

OTHER FW

- Plotter? Need to develop an efficient plotter code
  ↪At the moment using Antonio's code

- Fitting FW?

Today's Agenda

- Status of analysis: ntuples, preselection, trigger studies
- Status of ML
- Open discussion: analysis strategy, region definition, training…

Next Steps in the Analysis

- Analysis strategy
  - ↪"Freeze" trigger selection
  - ↪Define regions: SR, CR
  - ↪Background estimation
  - ↪W rediscovery
- ML strategy
  - ↪Training/validation region: where? How?
  - ↪Unblinding strategy design
- Uncertainties
  - ↪Signal?
  - ↪In ML?
- Manpower, timescale

Analysis Timescale

- What we need to converge?

- Manpower
  - ↪Antonio has 2 years
  - ↪Graziella has 3-4 months
  - ↪Michela?
  - ↪1/2 master students from Naples in the next months

Machine Learning

- Implement full machinery
- How do we create an event-level score?
- Train on MC QCD: how?

  ↪QCD in slice: select subsample? What about weights? —> Train for each slice

  - How evaluate score?

- Train on data: where?
- Uncertainties in ML