Efficient graph matching and association scores between reconstructed and simulated particle interactions

Bushra Almutrafi

bushra.almutrafi@kaust.edu.sa

Supervisors:

Prof. Felice Pantaleo
Dr. Elena Vernazza
Dr. Izaak Neutelings

Abstract

Comparing what a detector reconstructs to what actually happened in a simulated event is essential for judging how well reconstruction algorithms perform. This project develops a workflow to connect simulated ground-truth particles to the objects produced by reconstruction, using the CMS Truth Graph as the source of ground-truth information. A custom analysis tool was built to extract and validate this ground-truth data, confirming that particle counts, relationships, and decay classifications match known expectations, and that reconstructed detector energy closely tracks the true simulated energy. Truth particles and reconstructed objects were then matched using the detector hits they share, with a score quantifying the quality of each match. This matching was carried out in both directions — from truth particles to reconstructed objects, and from reconstructed objects back to truth particles — since checking only one direction can give an incomplete picture.

The results show that reconstructed objects are almost always genuine, clean matches to a single real particle, while real particles are recovered less consistently, particularly when they carry low energy. This asymmetry, visible only by comparing both directions, highlights where the reconstruction performs reliably and where it does not. The workflow developed here provides a systematic, repeatable way to evaluate reconstruction quality and can support future work on reconstruction performance studies and machine-learning-based reconstruction methods.

Date: August 24, 2026

Code: https://github.com/Bushra-Abdulrahman/Truth-Graph-With-Association

Contents

1 Introduction **2**

2 Background **2**

3 Methodology **3**

3.1 Dataset and Analysis Environment . . . 3

3.2 Data Extraction and Processing . . . 3

3.3 Data Validation . . . 3

3.4 Bidirectional Matching . . . 4

3.5 Evaluation Metrics . . . 6

4 Results **6**

4.1 Validation Results . . . 6

4.2 Truth → RECO Results . . . 7

4.3 RECO → Truth Results . . . 8

5 Discussion **9**

6 Conclusion **10**

References **11**

1 Introduction

In complex scientific detectors, reconstruction software tries to infer what happened during a particle collision starting from low-level sensor signals. At the same time, simulation provides a detailed ground-truth description of the event, including the particles that were generated and how they propagated through the detector. Comparing reconstructed information with this simulated reference is essential for evaluating reconstruction performance and identifying where reconstruction algorithms can be improved.

A central challenge is to match the reconstructed objects to the simulated one, in order to determine whether a reconstructed object corresponds to a true simulated particle and, if so, which one is the best match. Beyond simple matching, quantitative association scores are needed to evaluate the quality of these associations in both directions: from reconstructed objects to simulation and from simulation to reconstructed objects. This provides a systematic way to identify successful, missing, or ambiguous associations and evaluate the overall matching performance.

This project develops an analysis and validation workflow within the CMS software framework to study this problem. The workflow uses the Truth Graph [1] as the source of simulated truth information, validates the extracted information, performs bidirectional matching between selected truth and reconstructed objects, and evaluates the resulting associations using quantitative matching criteria.

2 Background

In the CMS experiment, when particles collide, they interact with the detector and leave signals that reconstruction algorithms use to identify the particles and estimate their properties. The output of this process is a set of reconstructed (RECO) objects.

To evaluate how well these algorithms perform, we need a reference to compare with. Simulation plays this role by providing a ground-truth description of what happened inside the detector. This information was previously stored in separate objects and collections, making it difficult to access and follow the connections between them. CMS now provides a new tool under development called the Truth Graph. It brings this information together into a unified representation, making truth information easier to access, follow, and analyze.

Once we have the reconstructed and truth information, we can establish associations between them to determine which reconstructed object corresponds to which truth particle. This association can be applied in both directions. In the Truth-to-RECO direction, we start from a truth particle and search for the reconstructed object associated with it. In the RECO-to-Truth direction, we start from a reconstructed object and search for the truth particle that most likely corresponds to it.

The association is also provide a score to quantify the quality of the match. This makes it possible to compare different candidate matches, select the best one, and evaluate how reliably the reconstructed and simulated information can be connected.

3 Methodology

3.1 Dataset and Analysis Environment

The project uses simulated CMS collision data containing both ground- truth and reconstructed information. *Simulated* means every event is generated by software, so unlike with real data, we already know exactly what particles were produced and how they should have behaved. This known answer is called the ground truth, and it is what makes it possible to check how well the reconstruction algorithms are doing: we can compare their output against something we already know is correct.

The main dataset used is a sample called TenTau, made of events that each contain ten simulated tau particles (a type of unstable particle that quickly decays into other, simpler particles). This sample is a convenient test case because the tau particle's behavior is well understood, which makes it easy to spot mistakes if the extracted data doesn't look right.

All of the work was done inside CMSSW [2], the software framework used by the CMS experiment to process and analyze this kind of data.

3.2 Data Extraction and Processing

The raw data is not directly usable, it needs to be read, filtered, and organized before it can be analyzed. This step is handled by a custom analysis tool built for this project. For every event, the tool reads the ground-truth information and the reconstructed information, pulls out the specific quantities needed (particle type, energy, relationships between particles, etc.), and saves them for later use.

An important part of the ground-truth information is that particles are connected to each other: each particle can have a parent (what produced it) and children (what it decayed into). The tool follows these connections, which makes it possible to select particles not only by type, but also by where they came from. for example, "a particle that came from a tau decay."

3.3 Data Validation

Before trusting the extracted ground-truth data enough to use it for matching, it needs to be checked for mistakes. If something in the extraction process is wrong, it usually shows up as an unexpected or unphysical result in a quantity we already know the correct answer for. The validation was performed at different levels, starting from individual particles and extending to more complex relationships and reconstructed information.

3.3.1 Particle Level Validation

The first stage checks basic properties of individual particles: counts per event, momentum and angular distributions, and parent/ancestor identity, for a few different particle types (electrons, photons, muons, taus). These are quantities with a known, expected behavior, so agreement between the extracted values and expectation is itself evidence that the graph is being read correctly. Ancestry relationships were also checked directly at this stage, by confirming that a particle's recorded origin (what produced it, and what it

decayed into) is consistent with the underlying physical process used to generate the sample.

3.3.2 Tau Decay Classification

The second stage is a classification task. Tau particles can decay in different ways, and each decay produces a different combination of final particles. For every tau particle, its decay products were followed and grouped into one of three categories. The resulting percentages were then compared against known reference values. A close match between the two confirms that the tool is correctly following particle relationships, not just reading single particles in isolation.

3.3.3 RECO-Truth Analysis

The next stage moves from particles to detector signals, and looks at how well the detector's reconstructed measurements match the original simulated ones.

When a particle passes through the detector, it deposits energy, and the detector records this as a "hit." Because this is simulated data, we have two versions of every hit: the *simulated hit*, which is the exact, true amount of energy the particle actually deposited, and the *reconstructed hit*, which is the value the detector reconstruction software estimates from the (also simulated) detector signal, the same kind of estimate it would produce on real data.

To compare the two, each simulated hit is linked to its matching reconstructed hit using a shared identifier that marks which part of the detector the hit occurred in. This link makes it possible to look up, for any particle, both the true energy it deposited and the energy the detector reconstruction recovered for the same location.

Using this link, a **detector response** was calculated for tau particles: the ratio of reconstructed energy to true energy.

$$\text{Energy Response} = \frac{\text{RECO Energy}}{\text{Truth Energy}}. \tag{1}$$

A response close to 1 means the reconstruction recovered almost all of the particle's energy; a lower response means some energy was lost or not fully captured. This response was also studied across different regions of the detector, to check whether the accuracy of the reconstruction changes depending on where in the detector the particle passed through.

This step matters because it validates the tool's ability to connect ground-truth particles to detector-level signals, which is the same basic mechanism used in the next stage to match truth particles to reconstructed objects.

3.4 Bidirectional Matching

After validating the extracted information, the next stage focuses on matching ground-truth particles with reconstructed objects. The matching is performed in both directions to evaluate the correspondence from the perspective of both simulation and reconstruction.

For each input object, the association algorithm provides a set of candidate matches together with an association score representing the quality of each match. The candidate with the best score is selected for further evaluation.

3.4.1 Truth → RECO Matching

Here the search starts from a truth object and looks for the reconstructed candidate that best represents it. As a working example, truth-level charged pions originating from decays were used: for each one, the algorithm collects the reconstructed Tracksters sharing hits with it, ranks them by score, and keeps the best match. This direction answers a recall-style question: of the truth particles we expect to find, how many were actually recovered by reconstruction, and how well?

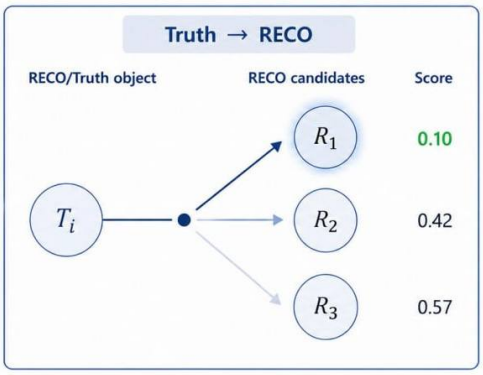


Figure 1 RECO to Truth Matching

3.4.2 RECO → Truth Matching

Here the search runs in reverse: starting from a reconstructed object, the algorithm looks for the truth candidate that best explains it, again ranking by association score and keeping the best match. This direction answers a precision-style question: of the objects the reconstruction produced, how many are cleanly explained by a single truth particle, rather than being a mix of several or dominated by noise?

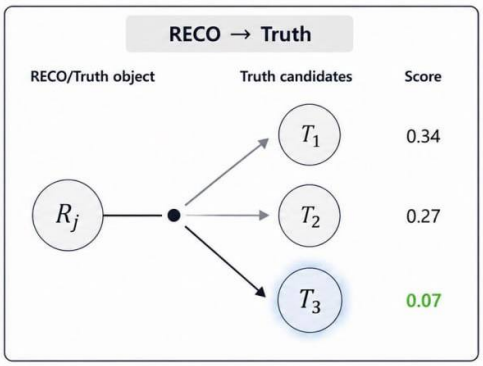


Figure 2 Truth to RECO Matching

Running both directions together, rather than only one, is what makes it possible to catch cases where a match looks good from one side but not from the other — for example, a truth particle whose best match is a given Trackster, while that same Trackster's own best match points to a different truth particle.

3.5 Evaluation Metrics

To turn the raw matching output into something that can be reported and compared, a threshold is applied to the association score to decide whether a match counts as "good": a score below 0.6 is treated as a successful match, following the convention that lower scores indicate better matches.

Using this threshold, a simple matching efficiency is defined for each direction:

$$\epsilon(p_T) = \frac{N(\text{objects with a best match with score} < 0.6)}{N(\text{all objects})}. \tag{2}$$

computed as a function of a chosen variable x (in this work, the transverse momentum, pT, of the truth or reconstructed object). In the Truth → RECO direction this efficiency behaves like a recall curve: it tracks how often truth particles are successfully recovered as a function of their pT.
In the RECO → Truth direction it behaves like a precision curve: it tracks how often reconstructed objects are cleanly attributable to a single truth particle. Reporting both gives a fuller picture of matching quality than either direction alone, and makes it possible to see whether performance depends on particle energy or detector region.

4 Results

4.1 Validation Results

The particle-level validation (Section 3.3.1) confirmed that basic quantities "how many particles of each type appear, their energy, and their parent particle" matched what was expected. No unusual or unphysical values were found, which gives confidence that the tool is reading and extracting the ground-truth data correctly.

The tau decay classification (Section 5.3.2) gave the following results, compared against the known reference values:

Table 1 Tau Decay Classification

Decay Mode	My Results (%)	Expected Values (%)
Muon	17.9	17.8
Electron	17.3	17.4
Hadronic	64.8	64.8

The extracted percentages are essentially identical to the expected values (within a fraction of a percent). This confirms that the tool correctly follows particle relationships

through the data, not just individual particles in isolation — which matters because the matching step later depends on the same kind of relationship-following.

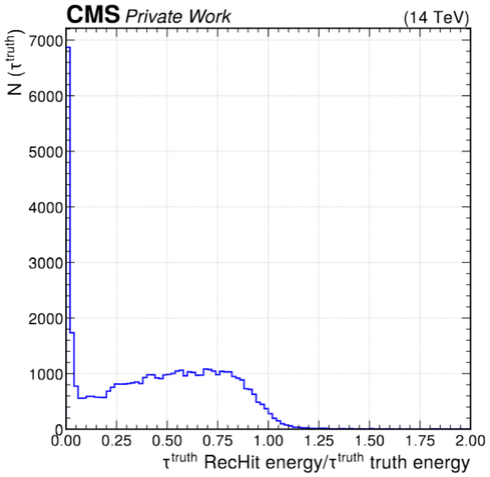


Figure 3 Energy Response of t

The RECO-Truth analysis (Section 3.3.3) compared reconstructed energy to true simulated energy for tau particles. The distribution has two distinct parts. A large spike sits right at 0: for a big share of tau particles, almost no reconstructed energy was recovered at all. This is expected rather than a mistake — it matches tau decays where the final particle (for example, a muon) passes through the detector without depositing much energy in the calorimeters, so there is very little for the reconstruction to pick up in the first place. Beyond this spike, the remaining particles form a broader, fairly flat distribution between about 0.5 and 1.0, meaning that when a particle does deposit a meaningful shower of energy, the reconstruction typically recovers a large majority of it. When this response is checked across different regions of the detector, it stays consistent fairly evenly, with no strong dependence on detector position. This confirms the reconstruction's energy accuracy is stable across the detector, not just correct in one region.

Together, these checks show that both the ground-truth data and its connection to detector hits can be trusted as a reliable reference for the matching results below.

4.2 Truth → RECO Results

For the Truth → RECO direction, the association-score distribution is heavily concentrated near the high end, close to 1.0, with comparatively few truth pions landing below the 0.6 cut. This means that, for a large share of truth pions, no strongly overlapping reconstructed object was found.

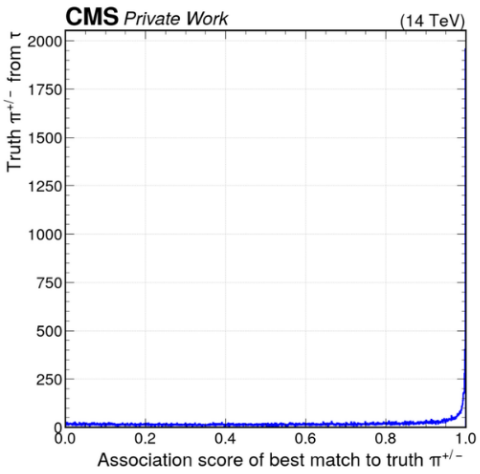


Figure 4 best score distribution

This is reflected directly in the matching efficiency as a function of truth pion energy efficiency starts near 0 for low-energy pions and rises steadily as energy increases, reaching roughly 60–80% in the higher-energy range. In other words, higher-energy particles are noticeably more likely to be picked up correctly by the reconstruction than lower-energy ones, which deposit less energy and leave a weaker, harder-to-identify signal in the detector.

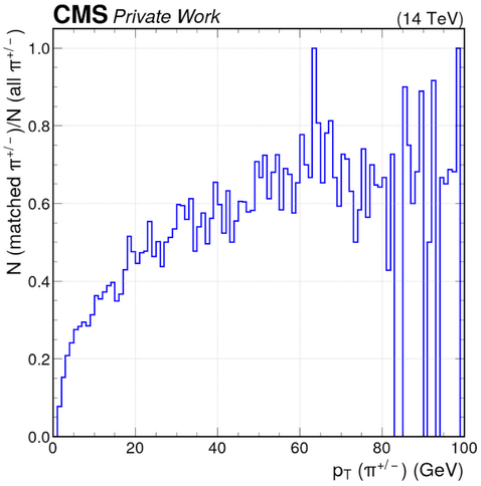


Figure 5 efficiency as a function of p_T

$$\epsilon(p_T) = \frac{N(\text{truth pions with a best match to trackster with score} < 0.6)}{N(\text{all truth pions})}. \tag{3}$$

4.3 RECO → Truth Results

The RECO → Truth direction shows a very different picture. The association-score distribution is sharply concentrated near 0, well below the 0.6 cut, meaning that the large majority of reconstructed objects have a strong, clean match to a single truth pion.

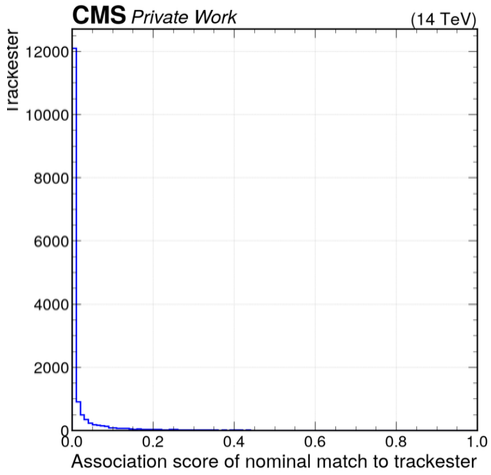


Figure 6 best score distribution

This is confirmed by the matching efficiency as a function of reconstructed object energy efficiency sits at or near 1.0 across almost the entire energy range, with only a few isolated dips at lower energy. In other words, nearly every reconstructed object corresponds cleanly to a real, simulated particle, regardless of its energy.

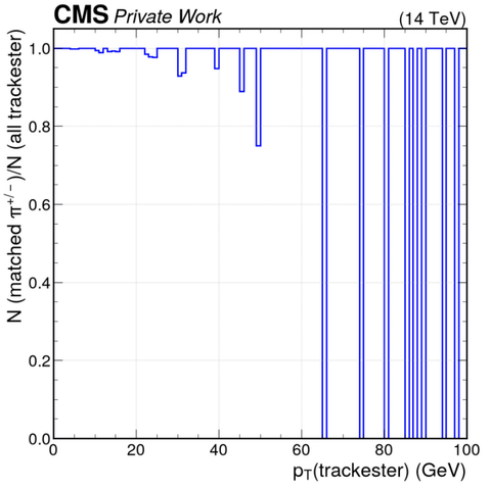


Figure 7 efficiency as a function of p_T

$$\epsilon(p_T) = \frac{N(\text{trackster with a best match to truth pion with score} < 0.6)}{N(\text{all trackster})}. \tag{4}$$

5 Discussion

The validation steps (Section 3.3) show that the ground-truth data extracted from the Truth Graph can be trusted: particle counts, relationships, and decay classifications all match known expectations closely. This matters because everything that follows the energy comparison and the matching itself depends on this data being correct. If the

extraction had errors, the matching results would not be meaningful, no matter how good the matching algorithm itself is.

The most notable finding is the clear difference between the two matching directions (Section 3.4). RECO → Truth matching is highly efficient and stays that way across almost the entire energy range: nearly every reconstructed object corresponds cleanly to a real, simulated particle. Truth → RECO matching is lower overall and clearly depends on energy: a real particle is more likely to be found by the reconstruction if it carries more energy, and less likely if it carries little.

This is not a contradiction between the two results, it reflects two different questions being asked. RECO → Truth asks "is this reconstructed object real?", and the answer here is mostly yes. Truth → RECO asks "was this real particle actually reconstructed?", and the answer depends on how much energy it had to begin with. Both are useful questions, and the value of running the matching in both directions is exactly that it separates these two, rather than folding them into a single number.

6 Conclusion

This project set out to build and test a way of connecting simulated ground-truth particles to the objects produced by detector reconstruction, and to measure the quality of that connection with a clear, repeatable score.

The work moved through several stages: understanding how truth and reconstructed data are represented, building a tool to extract and validate that data, studying how well reconstructed energy matches the true simulated energy, and finally matching truth particles to reconstructed objects in both directions and scoring the results.

The validation results confirm that the extracted ground-truth data is reliable. The matching results, computed in both directions, show a clear and consistent picture: reconstructed objects are almost always genuine, single-particle matches, while real particles are recovered less consistently, especially at lower energy. This shows that the overall approach is a workable and informative way to evaluate reconstruction, and that checking both directions gives a meaningfully more complete picture than checking only one.

References

[1] CMS Collaboration. *CMS Truth Documentation*. https://cms-truth.docs.cern.ch/. 2026.

[2] CMS Open Data. *CMS Open Data Guide: CMSSW Overview*. https://cms-opendata-guide.web.cern.ch/cmssw/cmsswoverview/. 2026.