Development of particle flow algorithm with GNN for Higgs factories

Tatsuki Murata[1,*] and *Taikan* Suehara[1,2,**]

[1]Graduate school of science, the University of Tokyo
[2]ICEPP, The University of Tokyo

Abstract. Particle flow plays an important role in precise measurement of Higgs bosons at future lepton colliders such as ILC and FCCee. Various detector concepts are designed to maximize the effect of particle flow to be able to separate each particles inside jets and improve the resolutions. For the standard particle flow algorithm, PandoraPFA is used for long in ILC studies. It is a multi-step reconstruction algorithm consisting of clustering, track-cluster association, and various refinement processes. We have studied machine learned particle flow model using Graph Neural Network based algorithm developed in the context of CMS HGCAL clustering. This model utilizes GravNet as GNN architecture and Object Condensation loss function for training. Since the HGCAL algorithm only performs clustering at the calorimeter, we have extended the model with track-cluster matching to achieve full PFA. Details of initial implementation of the track-cluster matching algorithm as well as performance evaluation with multiple tau events and jet events will be shown. The results are also compared to the Pandora PFA.

1 Introduction

1.1 Higgs factories

Higgs factories are the colliders that produce large quantity of the Higgs bosons. The purpose of the Higgs factories is to investigate the nature of the Higgs bosons, which still remains a mystery, and to verify physics beyond the Standard Model. The Higgs factory is primarily based on an electron-positron collider and is expected to operate at energy optimized for the efficient production of Higgs bosons.

A high resolution calorimeter is essential for the precise study of the Higgs boson, and the performance of the calorimeter is critical in the Higgs factory. Specifically, high jet energy resolution is a key factor that will significantly impact the quality of the scientific results. Jets need to be measured with high energy resolution to accurately reconstruct the decay products of the Higgs boson. For example, to correctly identify the Higgs boson decay modes, the jet energy resolution ideally needs to be within 3–4%.

*e-mail: murata@icepp.s.u-tokyo.ac.jp
**e-mail: suehara@icepp.s.u-tokyo.ac.jp

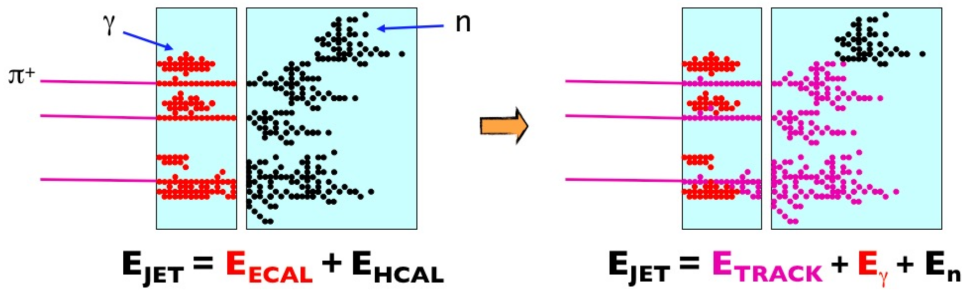

Figure 1. The concept of PFA. Charged particles (purple) are reconstructed and measured by the internal tracker, photons (red) by the ECAL, and neutral hadrons (black) by the HCAL [1]

1.2 Particle Flow Algorithm

In order to achieve the jet energy resolution of 3-4%, the Particle Flow Algorithm (PFA) has been proposed [1]. As shown in Figure 1 (left), the jet energy resolution was conventionally obtained as the sum of the measured energies in ECAL and HCAL. in a typical jet, about 72% of the energy (62% charged hadrons and 10% neutral hadrons) is measured by HCAL, and the typical resolution of HCAL of $55\%/\sqrt{E(\mathrm{GeV})}$ makes it difficult to achieve the target jet energy resolution. On the other hand, the PFA identifies each particle in the jet and determines its energy and momentum with the most appropriate detector, as shown in Figure 1 (right). In the HCAL, only long-lived neutral hadrons (about 10% of the jet energy) are measured. This allows the jet energy resolution to be significantly improved by measuring each particle with a detector having the optimum resolution.

The PFA reconstructs the particles in the jet from the charged tracks obtained by the tracking detector and the shower information obtained by the calorimeter. Particles bent by the magnetic field are reconstructed as charged particles by the internal tracker, and particles unaffected by the magnetic field are reconstructed as neutral particles by the calorimeter. Therefore, it is important to separate the charged particle showers from the neutral particle showers with high accuracy, and a highly granular calorimeter is required for this purpose.

The PandoraPFA [1][2] is the typical reconstruction program that incorporates PFA based on the ILD. It uses Monte Carlo simulations to reconstruct jets and verify the performance of the ILD. The PandoraPFA is composed of multiple complex processes, each of which contains parameters. These parameters must be appropriately tuned for each detector, but the parameters are manually determined, making adjustments challenging. There is potential to simplify the execution of the PFA by leveraging machine learning, which could automate and optimize this process.

2 Neural network

2.1 Network model

In recent years, deep learning models have become very popular in areas of high energy physics. One critical process is the reconstruction of hit information within detectors. Neural networks are employed to accurately reproduce the trajectories and energy distributions of particles within the detector. Therefore, we are establishing particle flow algorithm using neural network. The development of graph neural network (GNN) was undertaken by advancing the model developed by the CMS High Granularity Calorimeter group[3]. This model is consists of graph neural network, GravNet[4], and object condensation loss functions[5].

2.1.1 GravNet

In GravNet, each particle detector hit is treated as a vertex in a graph. The layer first projects these hits into an abstract space, creating coordinates for each hit. In this space, vertices (hits) are connected based on their Euclidean distance from one another. GravNet selects a fixed number of nearest neighbors for each vertex, typically using a Gaussian or exponential function to assign weights to these connections. The closer the neighboring vertices, the higher their influence on one another, hence the name "GravNet" – inspired by gravitational forces, where the influence decreases with distance.

The GravNet layer processes input data consisting of features for each hit (e.g., spatial coordinates, energy levels, and time stamps). The overall network network architecture is schematically represented in Figure 2. The core steps include:

1. Distance-Based Graph Construction: The network builds a graph by connecting each hit to its nearest neighbors in the learned representation space.

2. Feature Aggregation: Information is exchanged between connected hits. Each hit collects weighted feature information from its neighbors based on the distance, emphasizing contributions from nearby hits.

3. Feature Transformation: The aggregated features are transformed using dense layers, enabling the network to extract complex patterns in the data.

GravNet's ability to adaptively learn the spatial relationships between hits makes it particularly effective for handling sparse, irregular data, such as those encountered in calorimeters. It is not constrained by predefined detector geometry and can dynamically adjust to different types of detectors. This flexibility is crucial for experiments like those at the Large Hadron Collider (LHC), where detectors such as the CMS endcap calorimeter have irregular sensor arrangements. By efficiently clustering energy depositions from particle interactions, GravNet improves the accuracy and performance of particle identification tasks.

2.1.2 Object condensation

The training is optimized using the object condensation loss function which offers the possibility to handle inputs with variable length and to integrate regression and classification tasks via the minimisation of a common loss function. The central idea of object condensation is to enable a model to automatically group individual data points (e.g., detector hits) into distinct objects, all while predicting object properties like position, momentum, and class. This process is performed simultaneously without the need for multiple stages. The key to this method lies in using a learned clustering space, where points are assigned to "condensation points" that act as representatives of each object.

Two critical loss functions are used to guide the model during training to ensure accurate object clustering and reconstruction:

1. L_β (Condensation Loss) : This loss function is used to identify the key points of each object, known as condensation points. These condensation points serve as representative points for their respective objects and play a central role in object identification. The function guides the model to minimize the score for background or noise points while maximizing the score for points that belong to the core of an object.

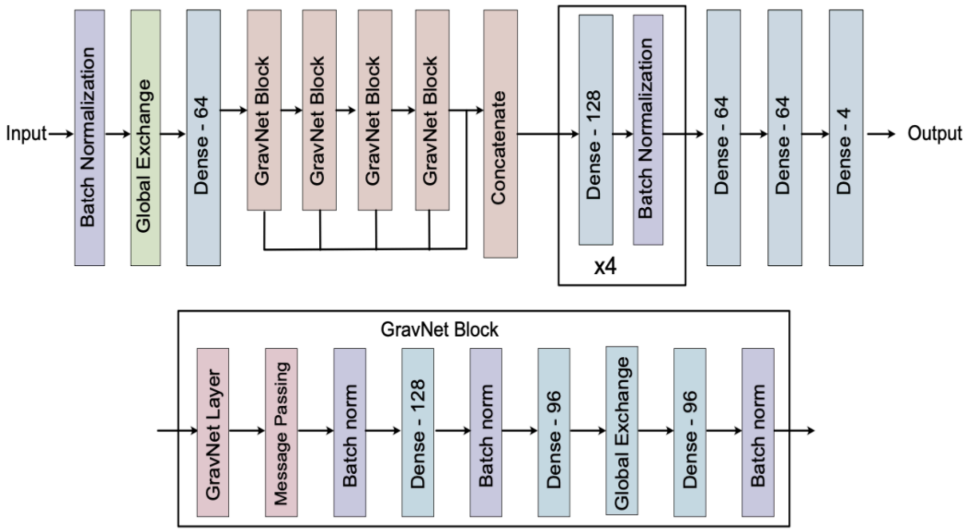

Figure 2. Architecture diagram of the GravNet graph neural network

2. L_V (Potential Loss) : This loss is based on potentials that pull data points together when they belong to the same object and push them apart when they belong to different objects. The attractive potential pulls points of the same object towards the condensation point, while the repulsive potential prevents overlap from other objects.

During training, the model learns to map data points into this latent space where clustering occurs. The L_β loss helps the network identify condensation points, while L_V ensures correct clustering. These two losses work together to allow the network to simultaneously learn both the structure of objects and their specific properties (like position, momentum, class). By allowing each data point to potentially represent an entire object, the method avoids the need for multi-stage algorithms.

In inference, condensation points are selected based on their β score. Starting with the highest β values, the model groups points around each condensation point based on their distances in the latent space. This approach efficiently reconstructs objects without predefined boundaries or sizes, making it robust for particle detection tasks where multiple, often overlapping particles need to be identified.

CMS HGCAL group has used this algorithm solely with the calorimeters; however, to implement PFA, tracker information is also required. To incorporate tracker data, "virtual" hits are assigned at the entry points on the calorimeter, extrapolated from the tracker trajectories. For the MC truth clusters of charged particles that leaves tracks in the tracker, the virtual hits are forcibly treated as condensation points regardless of the value of β by modifying loss functions. This also has the effect of bringing the β of tracker points closer to the maximum value of 1.

2.2 clustering method

The GravNet GNN described in 2.1.1 only gives the feature vectors and spatial coordinate of the vertices. therefore, another clustering algorithm needs to be done to perform particle flow. It is possible to implement other neural networks to perform clustering of hits or predict cluster energy. However, since the outputs of GravNet GNN includes spatial coordinates and the hits belonging to the same cluster would be located closer and other hits are spatially separated due to the characteristics of the object condensation loss function, we adopted distance based clustering algorithm. The algorithm consists of the following processes:

1. Points with beta above a certain threshold t_β ($\beta > t_\beta$) are defined as condensation points

2. The remaining points are classified as belonging to the nearest condensation point if the distance between the point and the condensation point is below a certain distance t_d.

3. Among the remaining unclustered points, point with the highest beta is assumed as the condensation point and cluster the unclustered point within the distance of t_d.

4. Repeat 3 until all points are clustered.

3 Dataset

ILD full detector simulation samples are used to train the model and evaluate the results.

In order to demonstrate the performance of our model, we produced a data set with ten τ^- particles per event with the energy of 10 GeV emanating from the interaction point at the center of the detector in random directions. The reason τ samples are used is because τ has a good mixture of decay product (i.e. pion, electron, muon, and photons) and clearer environment compared to quark jets which makes training of GNN easier. We generated 100,000 events and 80% of these was used for training GNN, 10% are used for validation, and the remaining 10% was used to evaluate the performance. We also generated $u\bar{u}, d\bar{d}, s\bar{s}$ jet events with 91 GeV, which are the official sample for PFA calibrations to compare the performance with the Pandora PFA. The quark jets are also emanated from the interaction point at the center of the detector in random directions. We generated 75,000 events and 80% of these was used for training GNN, 10% are used for validation, and the remaining 10% was used to evaluate the performance. The performance of GNN are evaluated from clustering capability of the decay products of generated samples, such as electron, pion, and photons.

4 Clustering performance

4.1 Quantitative evaluation

To evaluate the GNN clustering performance, $efficiency$ and $purity$ are calculated for all reconstructed clusters.

$$\text{Efficiency} = \frac{E_{\text{match}}}{E_{\text{MC}}}$$

$$\text{Purity} = \frac{E_{\text{match}}}{E_{\text{reco}}}$$

where $E_{MC}, E_{reco}, E_{match}$ represent MC truth cluster energy, reconstructed cluster energy, and correctly clustered energy respectively. Here, all cluster energies are defined as the sum of the energy deposits at each hit, with no corrections applied for factors such as the thickness of the absorber layers. This efficiency and purity serve as indicators of how accurately GNN

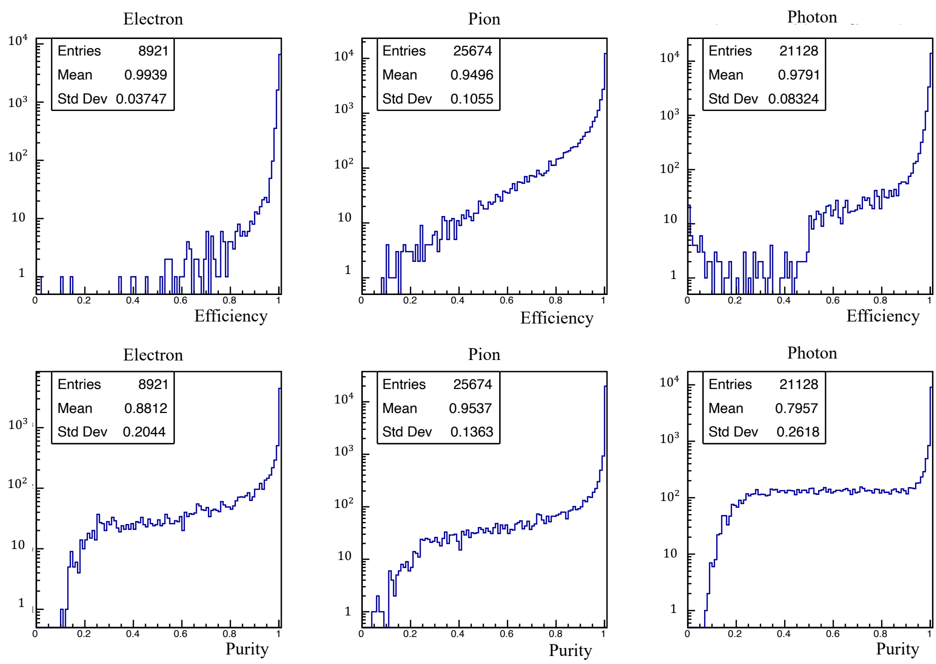

Figure 3. Efficiency (top) and purity (bottom) of $e^{\pm}$, $\pi^{\pm}$, and γ clusters from tau samples

correctly cluster hits, with values closer to 1 indicating better clustering performance. Efficiency indicates how much correct hits have been collected and purity indicated how many incorrect hits have not been included to reconstructed clusters.

Figure 3 shows results of efficiency and purity calculated from tau samples. The histograms are drawn by each decay products ($e^{\pm}$, $\pi^{\pm}$, and γ). For all histograms, one entry represents one MC truth cluster. In these histograms, only clusters with $E_{MC} > 1$ GeV are included. This is because for the clusters with MC energy below 1 GeV, clustering tend to bad due to fewer number of hits, and those particles would rarely reach calorimeter in ILD due to the inner radius of the calorimeter and magnitude of bending magnet. A pronounced peak is observed near 1 for all types of decay particles, suggesting that the clustering is performed successfully.

4.2 Hyperparameter tuning

This GNN method involves multiple hyperparameters: output dimensions (D), threshold of β to make condensation points (t_β), distance in virtual coordinate used for clustering (t_d). The output dimension (D) is the hyperparameter of the neural network which represent virtual coordinate of nodes of hits in GNN. To tune the parameter D, we trained the model with output dimensions of 2, 3, 4, 8, and 16, selecting the parameter that resulted in the highest efficiency and purity. β and t_d are the hyperparameters of the clustering algorithm we are adopting using GNN outputs written in 2.2. We performed a grid scan of both β and D in increments of 0.1 from 0 to 1, selecting the values that simultaneously maximized efficiency and purity.

4.3 Comparison with Pandora PFA

The result of GNN is shown in Table 1 after adjusting hyperparameters described in 4.2. The values presented in this table represent the mean of the histograms for each decay particle and each algorithm like histograms in Figure-3. The results from the GNN outperform those from PandoraPFA in most cases. The τ^- sample environment is cleaner with fewer number of decay particle and less background, leading to improved performance compared to the jet sample results. In most results, the performance exceeds 90%, but the electron purity is lower, around 77%. This reduction is likely due to the difficulty in separating electron-induced hits from photon-induced hits generated by bremsstrahlung, as they tend to be located in close proximity within the calorimeter. It should also be noted that these metrics reflect how accurately the hit information can be clustered, and a direct comparison with the PandoraPFA, which is specifically designed to optimize jet energy resolution, may not be entirely fair.

Table 1. Efficiency and purity of reconstructed clusters categorized by each decay product for τ^- and jets samples for GNN and Pandora PFA

		efficiency [%]			purity [%]		
sample	algorithm	$e^\pm$	$\pi^\pm$	γ	$e^\pm$	$\pi^\pm$	γ
τ^-	GNN	98.8	99.6	99.1	92.6	99.3	97.7
	Pandora PFA	99.3	94.0	99.1	91.8	94.6	97.2
jets	GNN	94.6	93.1	95.2	77.4	93.1	92.4
	Pandora PFA	80.2	90.4	79.0	75.0	90.6	77.7

5 Conclusions

We are developing a Particle Flow Algorithm using machine learning techniques. We start from a model developed by the CMS HGCAL group, which utilizing GravNet as GNN architecture and Object Condensation loss functions for training. Since the model only focuses on in using calorimeter hits, we implemented to add tracker information by adding virtual hits at the incident position in the calorimeter with no energy deposition and force them to be condensation points of the cluster. By applying this model to the ILD full detector simulation and implementing track-cluster matching algorithm, the performance of GNN based PFA was evaluated with multiple tau events and jets events. The performances are evaluated by two values, $efficiency$ and *purity*, both of which represent the accuracy of clustering. GNN based PFA has demonstrated improved clustering performance compared to currently widely used algorithm, Pandora PFA, through hyperparameter tuning. For the further comparison with PandoraPFA in terms of jet energy resolution remains for the future prospects.

We developed a Particle Flow Algorithm using Graph Neural Networks, based on the CMS HGCAL model. By incorporating track-cluster matching and applying this model to ILD simulation, we demonstrated improved clustering performance compared to the widely-used PandoraPFA. Future work will focus on further optimization of jet energy resolution and expanding the applicability of the GNN-based PFA to a broader range of events.

References

[1] J. S. Marshall, et al. "Pandora Particle Flow Algorithm", arXiv:1308.4537 [physics.ins-det]

[2] M.A. Thomson, "Particle Flow Calorimetry and the PandoraPFA Algorithm," Nucl. Instrum. Meth. A, 611, 25–40, 2009

[3] S. R. Qasim, et al. "Multi-particle reconstruction in the High Granularity Calorimeter using object condensation and graph neural networks", arXiv: 2106.01832 [physics.ins-det]

[4] S. R. Qasim, et al. "Leaning representations of irregular particle-detector geometry with distance weighted graph neural networks", arXiv: 1902.07987 [physics.data-an]

[5] J. Kieseler. "Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph and image data", arXiv: 2002.03605 [physics.data-an]