東京大学
素粒子物理国際研究センター
International Center for Elementary Particle Physics
The University of Tokyo

Study on High-Precision Jet Flavor Tagging at the ILC Using ParT (Particle Transformer)

Takahiro Kawahara[A], Taikan Suehara[B], Tomohiko Tanabe[C], Risako Tagami[A]

Sci. UTokyo[A], ICEPP UTokyo[B], MI-6[C]

Table of Contents

Table of Contents

Next Generation Higgs Factory

・Precision measurement of the Higgs boson is essential for exploring physics beyond the Standard Model.

➡ The construction of a Higgs factory is a strategic priority for the global particle physics community.

・Several designs for electron-positron colliders are being actively considered.

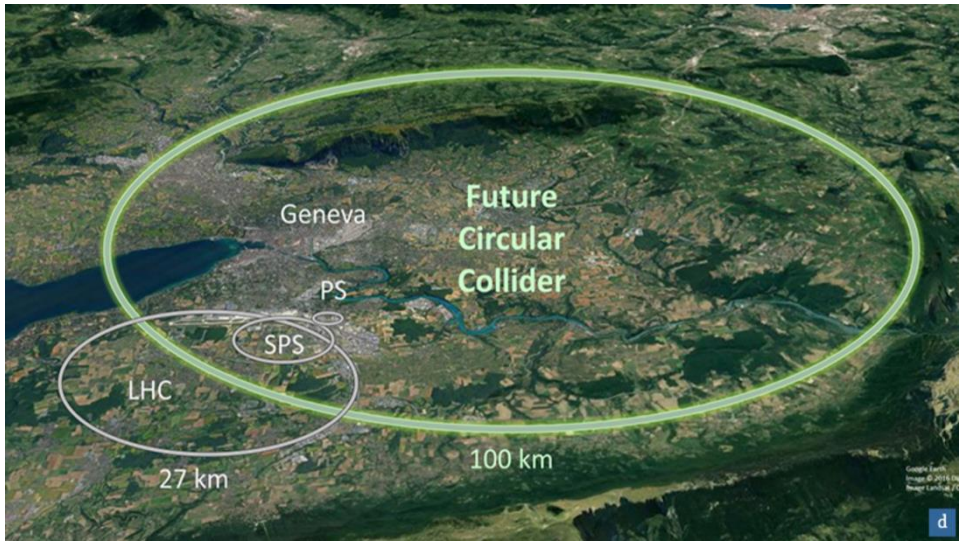
- **ILC (Japan) -** Linear Collider
  - **Detectors：ILD、SiD**

- **FCCee (CERN), CEPC(China) -** Circular Colliders

- Others in conceptual design phases…

ILC

FCC



ILD detector

The ILD detector consists of the following components, arranged from the inside out:

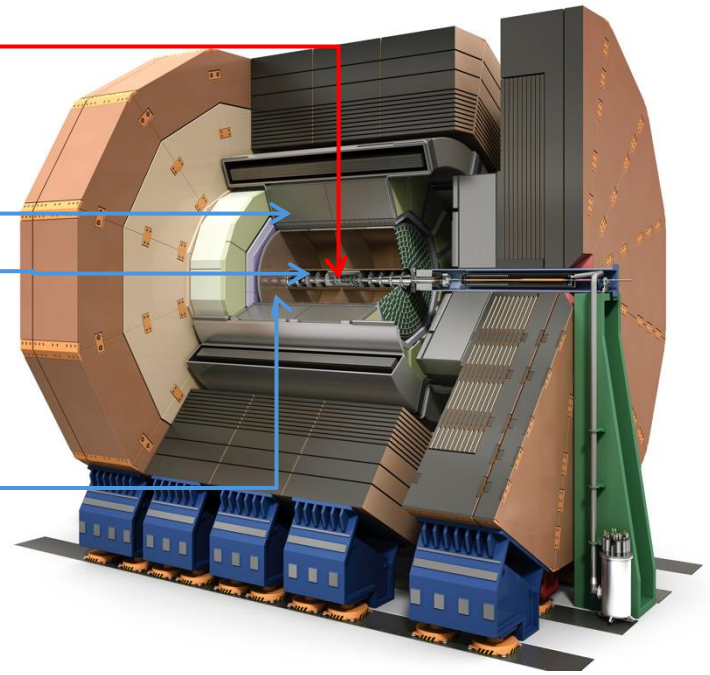
Note: Since impact parameters are critical for this study (b/c-tagging), this component requires the highest precision.

–**Vertex Detector**:
Accurately reconstructs the motion of charged particles near the interaction point.

–**Tracker**:
Precisely measures the momentum of charged particles.

–**Electromagnetic Calorimeter (ECAL)**:
Accurately determines the energy and position of photons.

–**Hadron Calorimeter (HCAL)**:
Measures the position, direction, and energy of hadrons.

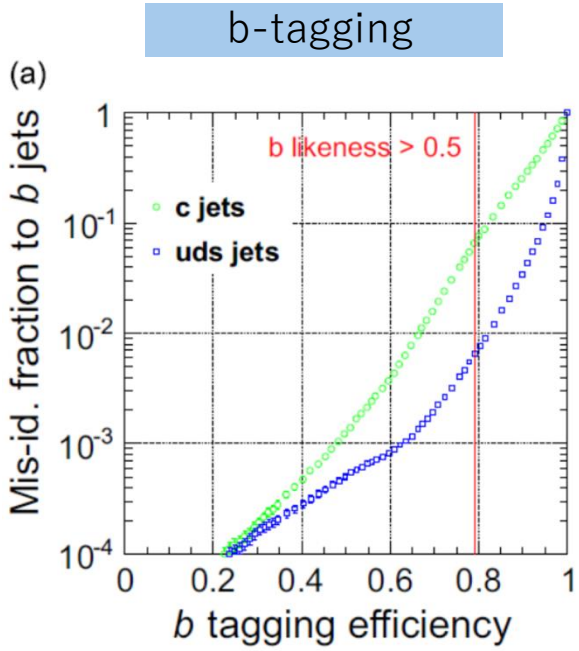


Flavor tagging for Higgs factories

Precise measurements of coupling constants such as **H→bb, cc, gg, ss** are crucial for exploring new physics. Therefore, improving the performance of **flavor tagging** is essential to achieve this goal.

- The software **LCFIPlus** (published in 2013) has been used for flavor tagging in ILC and CLIC research.
  - **Method**: Flavor tagging using Machine Learning (**BDT**).
- Recently, the **FCC-ee** group reported performance improvements of approximately **10 times** (for b- and c-tagging).
  - **Method**: Flavor tagging using **ParticleNet (GNN)**.
    - Note: Only fast simulation data was used.
- Currently, research utilizing the **Particle Transformer (ParT)** is being conducted by the ILC group, concurrently with the LHC group.

➡ In this study, I evaluated future possibilities using a model trained on the SGV (high-speed simulation) dataset employing ParT.



LCFIPlus performane plots

Table of Contents

Particle Transformer (ParT)

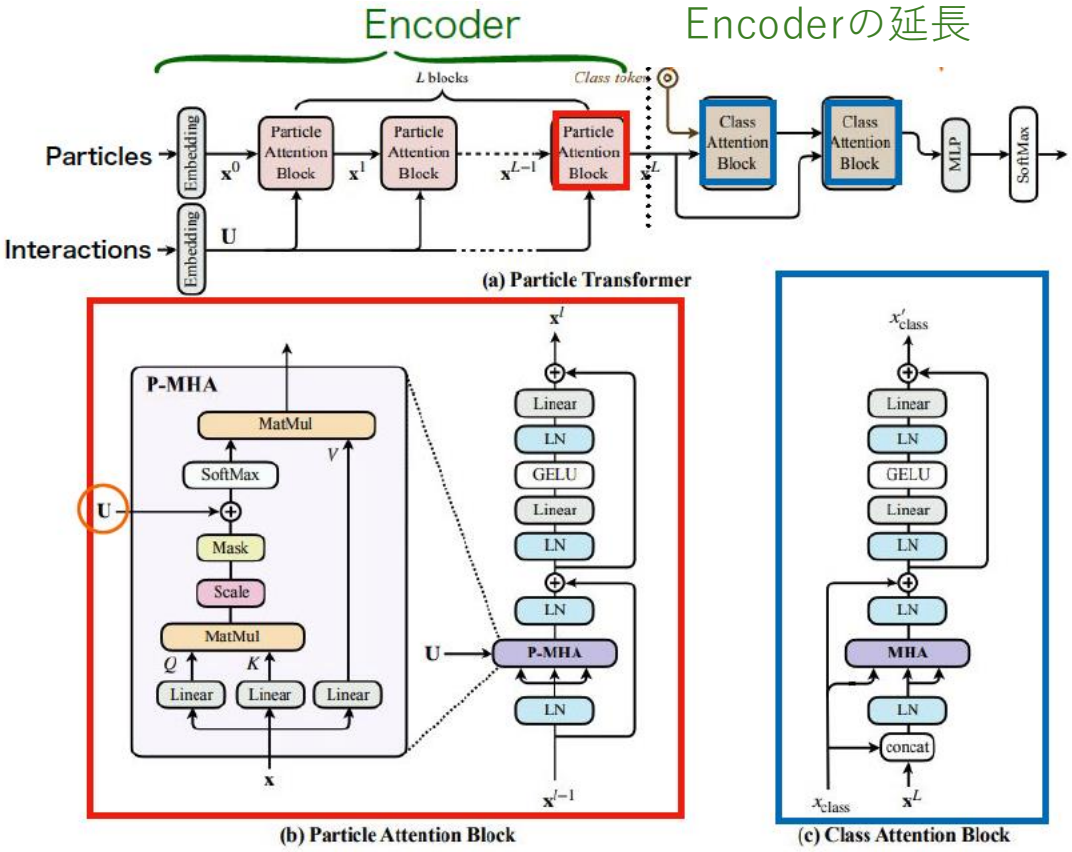
・**Transformer:** An algorithm based on self-attention. It is widely used in natural language processing (e.g., ChatGPT).

・**ParT:** A novel Transformer-based architecture published in 2022 for jet tagging.

- Enhances the attention mechanism by adding pair variables (angle, mass, etc.) to the standard Transformer encoder.

・Outperforms ParticleNet.

-While ParticleNet can only see "neighboring" particles, Transformers learn "where to pay attention" using the attention mechanism.



10.48550/arXiv.2408.12377

Basic Network Settings

- **Input:** Features of charged + neutral particles, Pairwise features between particles (4 dimensions).

- **Hyperparameters**:
  - Encoder: 8 Particle Attention Block,
  - Embedding dimension=128,
  - Head=8,
  - Decoder: 2 Class Attention Block,
  - **Activation Function:** GELU,
  - **Dropout:** None,

- **Output:** Flavor classification via softmax (CrossEntropy loss), representing the likelihood of being a b-jet.

Training Settings

Item	Setting
Batch size	256
Optimizer	AdamW
LR schedule	Cosine scheduler
Epochs	20
Loss	CrossEntropy
Hardware	GPU
Learning rate	Weight update step size

Dataset

- SGV (Simplified Generator-based Simulation)

  - $e^+e^- \rightarrow \nu\nu H \rightarrow \nu\nu qq$ (at 250 GeV)

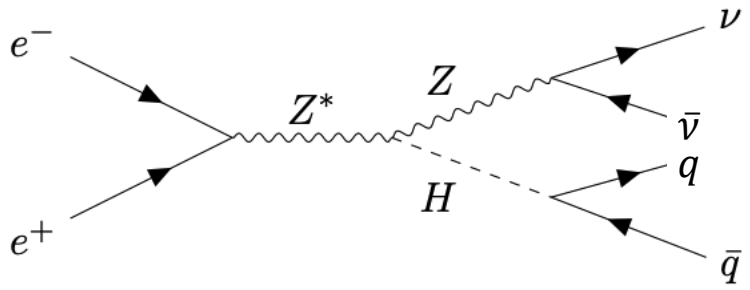
  - Total Events: 500K, 1M, 2.5M, 5M (train/val/test = 80:5:15%)
    - After splitting events, clustering is performed. Each resulting jet is treated as an independent sample (Inputs: 1M jets, 2M jets, 5M jets etc.).

- Comparison Dataset

  - ILD full simulation

  - Total Events: ~1M (Split: 80:5:15)
    - After splitting events, clustering is performed. Each resulting jet is treated as an independent sample (Inputs: 1M jets).

Higgsstrahlung Process

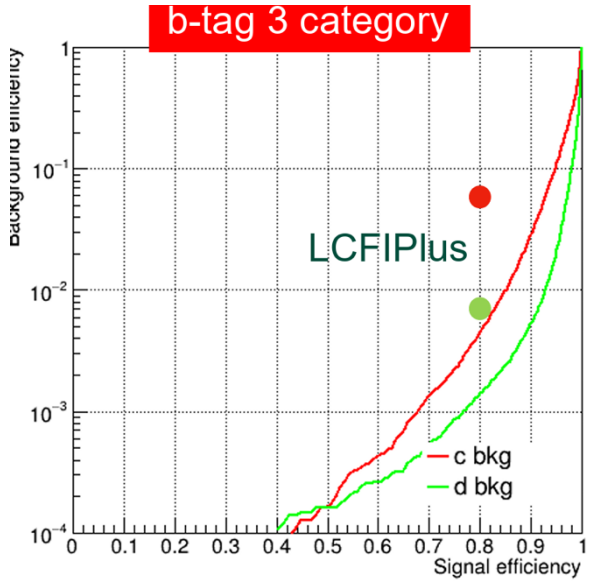


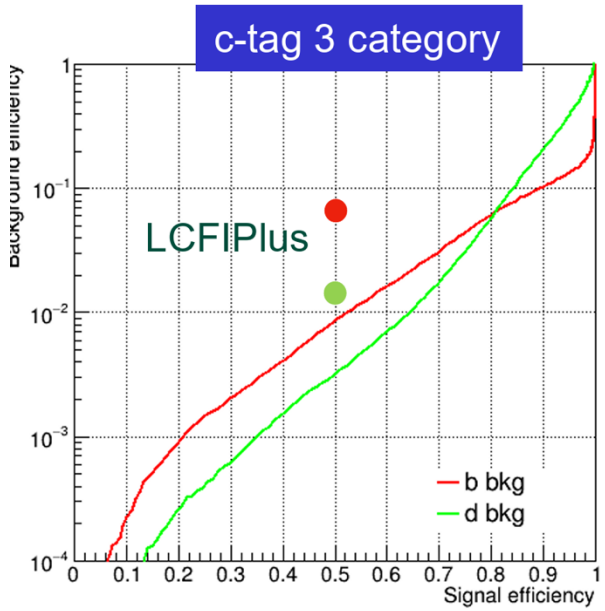
q = b,c,d
ν = neutrino

*As this study focuses on b-tagging, the background events are c and d jets.

Table of Contents

Comparison of LCFIPlus and ParT (ILD full simulation)





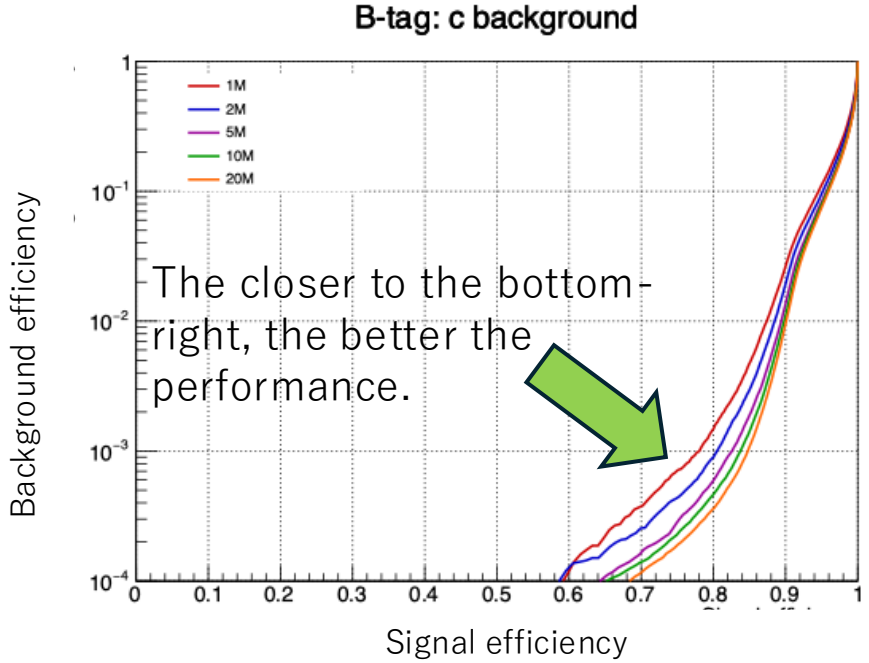
ParT has dramatically improved from LCFIPlus.

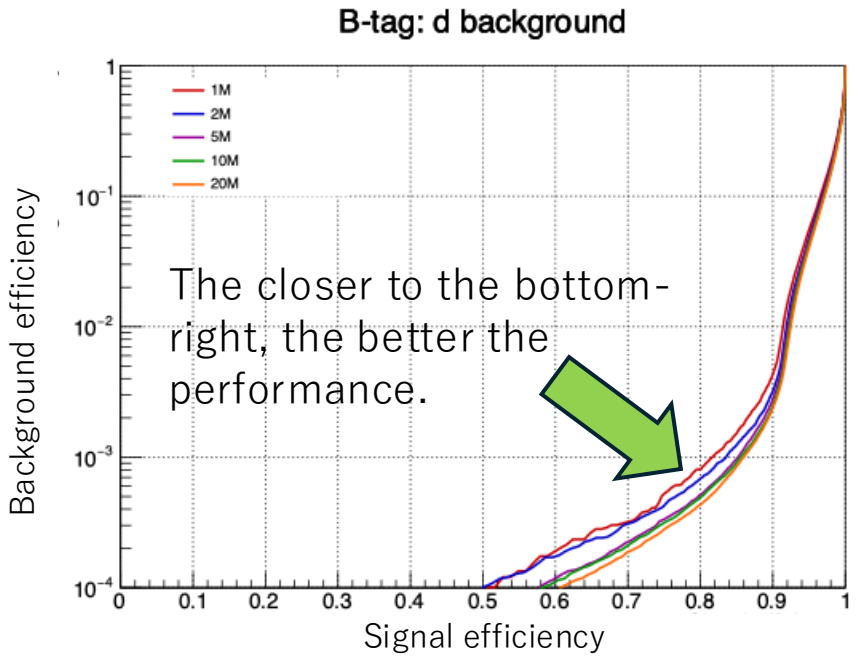
	b-tag 80% eff.		c-tag 80% eff.	
background	c jets	uds jets	b jets	uds jets
LCFIPlus	6.3%	0.79%	7.4%	1.2%
ParT	0.48%	0.14%	0.86%	0.34%

Approximately 13 times ！ ！

SGV Scalling Laws (Fixed at parameter count of 2 million)

B-tag performance, 3category　(Trained for 20 epochs on 1M, 2M, 5M, 10M and 20Mdatasets.)



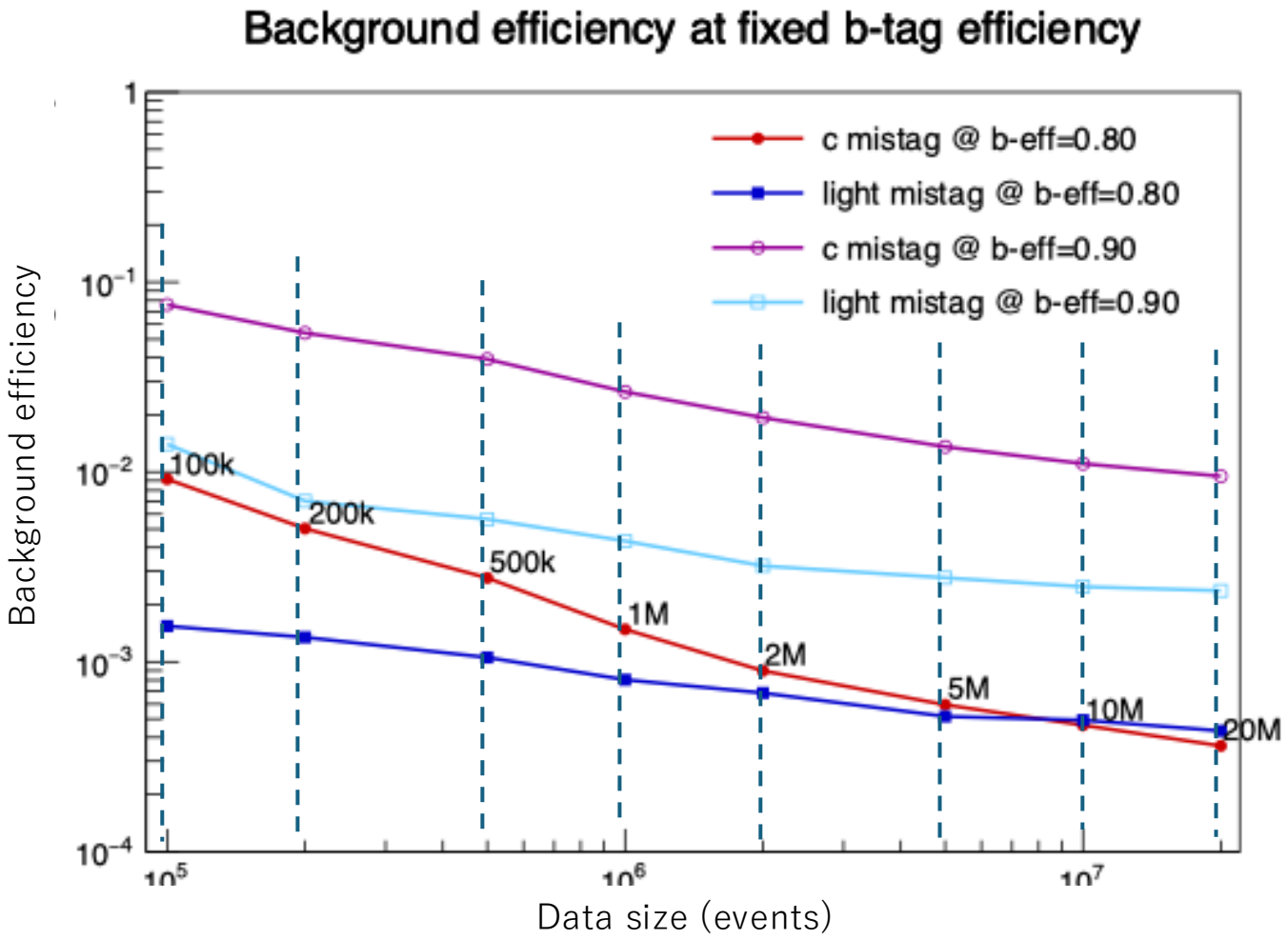


＊The horizontal axis shows signal efficiency (the percentage of b jets correctly identified as b), while the vertical axis shows background efficiency (the percentage of d and c jets mistakenly identified as b). The curves compare performance for different numbers of events used for training (1M, 2M, 5M, 10M).

b-tag 80% eff.	c bkg.	d bkg.
Fast sim (SGV) 1M	0.1480%	0.0805%
Fast sim (SGV) 2M	0.0897%	0.0684%
Fast sim (SGV) 5M	0.0597%	0.0516%
Fast sim (SGV) 10M	0.0464%	0.0493%
Fast sim (SGV) 20M	0.0331%	0.0286%

b-tag 90% eff.	c bkg.	d bkg.
Fast sim (SGV) 1M	2.645%	0.433%
Fast sim (SGV) 2M	1.931%	0.320%
Fast sim (SGV) 5M	1.354%	0.276%
Fast sim (SGV) 10M	1.102%	0.249%
Fast sim (SGV) 20M	1.005%	0.213%

SGV Performance Evaluation (Fixed at parameter count of 2 million)

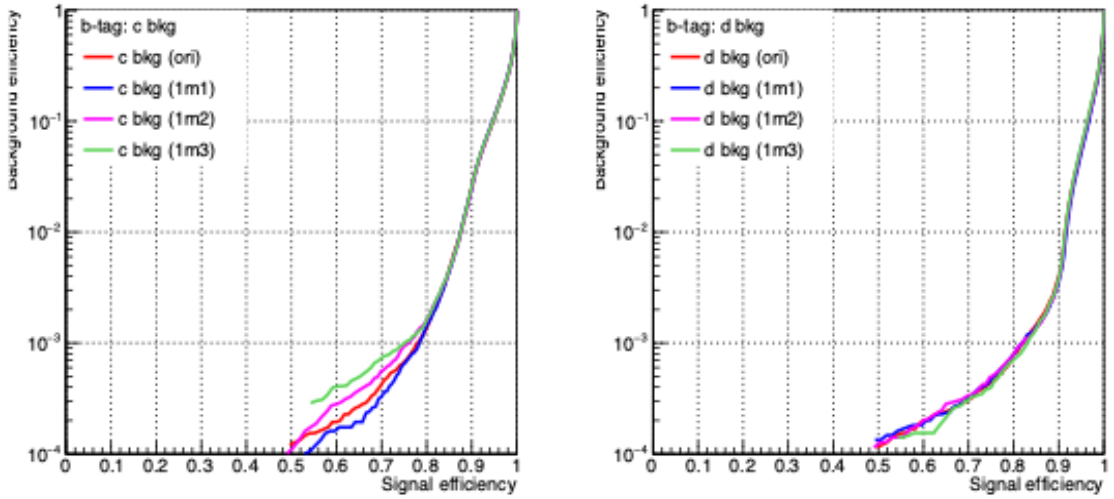


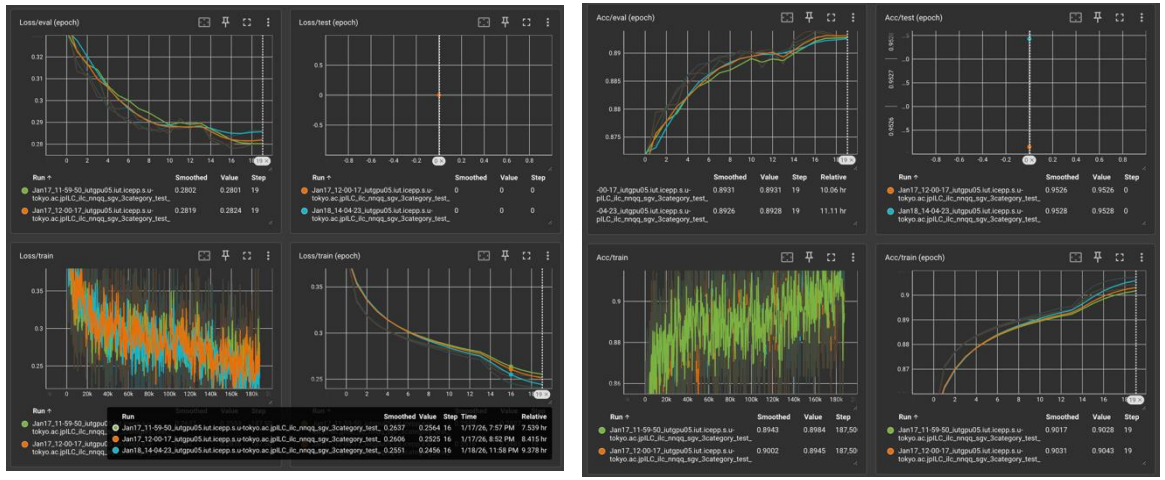
The figure on the left shows that as training data increases from 1M to 20M, the background efficiency for c and light jets decreases steadily. This corresponds to the "scaling law," where model performance improves with increased data size.

We plan to investigate cases where network complexity (number of parameters) is varied in the future.

SGV Performance Evaluation (Fixed at 1 million data)

Verified performance limits by scaling up the **Model Size (Width & Depth)** with a fixed dataset (1M).



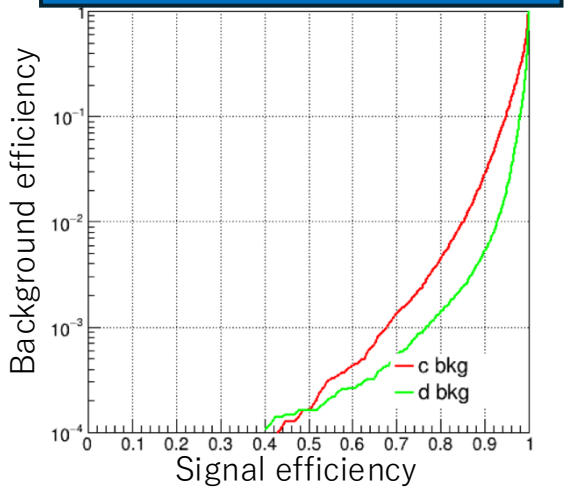


- **Configurations (Maintained Aspect Ratio):**
  - **Baseline (`ori`):** Layer 3 / Dim 128 / Heads 8
  - **Scale 1 (`1m1`):** Layer 4 / Dim 160 / Heads 10
  - **Scale 2 (`1m2`):** Layer 5 / Dim 192 / Heads 12
  - **Scale 3 (`1m3`):** Layer 6 / Dim 256 / Heads 16 (Large Scale)

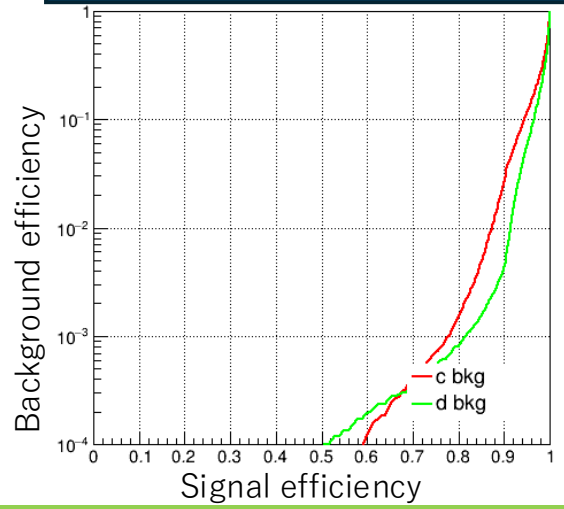
- **Observation:**
  - **Saturation:** Performance saturated around Scale 1 (Layer 4).
  - **Overfitting:** The largest model (`1m3`, Green line) shows clear degradation in ROC, indicating **overfitting**.
- **Conclusion:** The 1M dataset is insufficient for larger-scale models (wider & deeper). I will increase the dataset size to verify how this performance curve extends with more data.

Comparison: Fullsim vs SGV (at 1M)

Fullsim



SGV

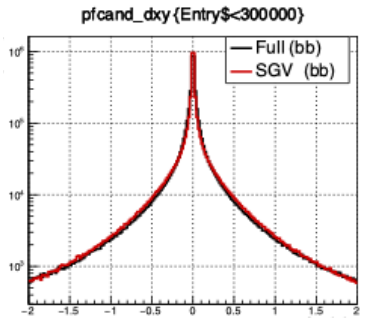


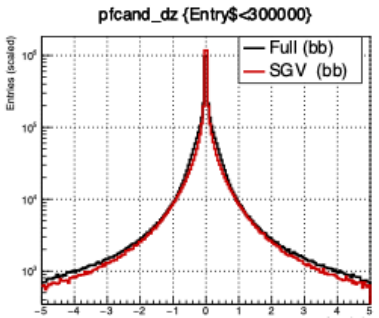
Key Results

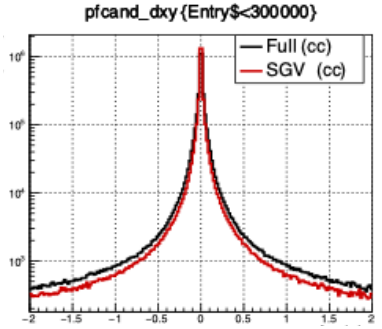
- A difference was observed between SGV and Full Sim. We compared input variable distributions to investigate the cause.

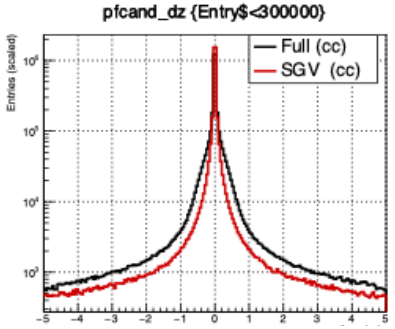
b-tag 80% eff.	c bkg.	d bkg.
Full sim 1M	0.627%	0.106%
Fast sim (SGV) 1M	0.148%	0.0805%

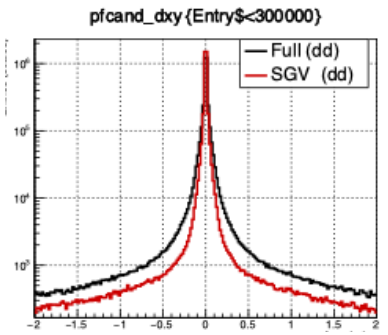
~Impact parameters~

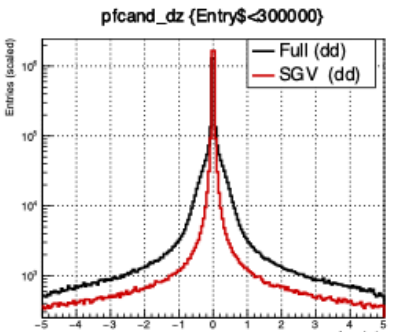












Summary

・ **Flavor tagging** is crucial for the search for new physics through precision Higgs measurements, and performance improvements driven by machine learning can contribute to the reach of these studies.

・ I introduced the **Particle Transformer** and achieved performance in b/c tagging that significantly exceeds conventional methods.

・ I am verifying scaling laws using **Fast simulation (SGV)**. Significant performance improvements are observed increases in both the number of samples and the number of parameters..

・**Future Plans**

- Verify scaling laws with even larger statistics and Full simulation to determine performance limits. (Data preparation in progress)
- Verify the effectiveness of pre-training with **SGV** (followed by fine-tuning with **Full simulation**).
- Investigate SGV tuning and correction methods using Full simulation.
- Scrutinize input variables and hyperparameters to aim for further performance improvements.
- Increase the amount of data to investigate at what stage performance saturation occurs.

Back up

Input Variable - Features

Impact parameter (charged) (6)
- pfcand_dxy,
- pfcand_dz,
- pfcand_btagSip2dVal,
- pfcand_btagSip2dSig
- pfcand_btagSip3dVal,
- pfcand_btagSip3dSig

Jet Distance (charged) (2)
- pfcand_btagJetDistVal,
- pfcand_btagJetDistSig

Particle ID flags (charged) (6)
- pfcand_isMu,
- pfcand_isEl,
- pfcand_isGamma
- pfcand_isChargedHad,
- pfcand_isNeutralHad
- pfcand_type

Track kinematics (charged) (4)
- pfcand_erel_log
- pfcand_thetarel,
- pfcand_phirel
- pfcand_charge

Track error / charged (15)
- pfcand_dptdpt,
- pfcand_detadeta,
- pfcand_dphidphi
- pfcand_dxydxy,
- pfcand_dzdz,
- pfcand_dxydz,
- pfcand_dphidxy
- pfcand_dlambdadz
- pfcand_dxyc,
- pfcand_dxyctgtheta
- pfcand_phic,
- pfcand_phidz,
- pfcand_phictgtheta
- pfcand_cdz,
- pfcand_cctgtheta

Network Settings Details

- Network Detailed Structure
  - **embed_dims = [128, 128, 128]**
    → Embedding dimension for each hidden layer
  - **pair_embed_dims = [64, 64, 64]**
    → Dimensions of the layer transforming pair features
  - **num_heads = 8**
    → Number of heads for Multi-head Attention
  - **num_layers = 3**
    → Number of layers for the Particle Attention Block
  - **num_cls_layers = 1**
    → Number of layers for the Class Attention Block (decoder side)
  - **activation = 'gelu'**
    → Activation function

About Fullsim and SGV

	Fullsim (GEANT4)	SGV (FastSim)
Accuracy	High	Approximate
Computational Cost	Very high	Low (mass generation possible)
Purpose	Final precision evaluation	Scaling laws and large-scale learning

- **Generate large-scale samples using SGV → Investigate scaling laws (performance vs. data volume)**
- **Compare FullSim/SGV → Verify consistency of variable distributions to ensure SGV reliability**
- **Future applications**
  - Pre-training with SGV
  - Fine-tuning with FullSim

  → Leverage the strengths of both approaches

⇨ 「SGV is essential for large-scale analysis, but demonstrating its reliability through comparison with FullSim is crucial.」
「The combination of SGV and FullSim enables efficient and precise analysis.」

Data preprocessing

- Data Partitioning

For a total of 5 million events:
After partitioning events, cluster each event and input the resulting jets as independent samples into 10Mjet.
Train : Val : Test = 8M : 0.5M : 1.5M (80:5:15)
Range partitioning by entry → Ensures reproducibility and independence

- Common Pre-Cut

Set physical upper limits on momentum/energy for charged and neutral particles
Exclude non-physical events and outliers, maintaining consistency across all regions

- Class-Specific Data

Process the 3 categories (b, c, d) independently
Apply identical conditions within each category

Reconstruction of Events

Tracking
Reconstructing charged particle tracks from detector hits
Extracting momentum and impact parameters through fitting
Neutral particles are measured by the calorimeter (ECAL/HCAL)

Collapse Point Detection (Vertexing)
-Primary Vertex (IP): Vertex generated at the collision point
-Secondary Vertex: Collapse point of quasi-stable particles (B, D hadrons, etc.)
Fitting using multiple tracks to estimate the collapse point

Jet Clustering
Clustering particles using algorithms such as Durham
Incorporating decay point information to construct more precise jet structures

Flavor Tagging
Input decay points, track characteristics, and energy distribution
Identify b, c, light quark/gluon jets