Application of Particle Transformer for Quark Flavor Tagging on Future Higgs Factories

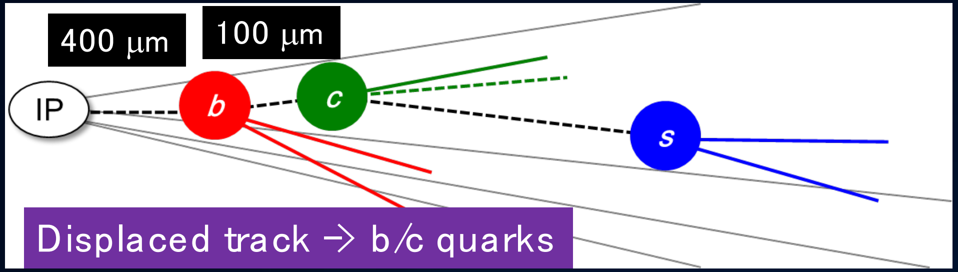
Presenter: Lai Gui

Instructor: Dr. Taikan Suehara

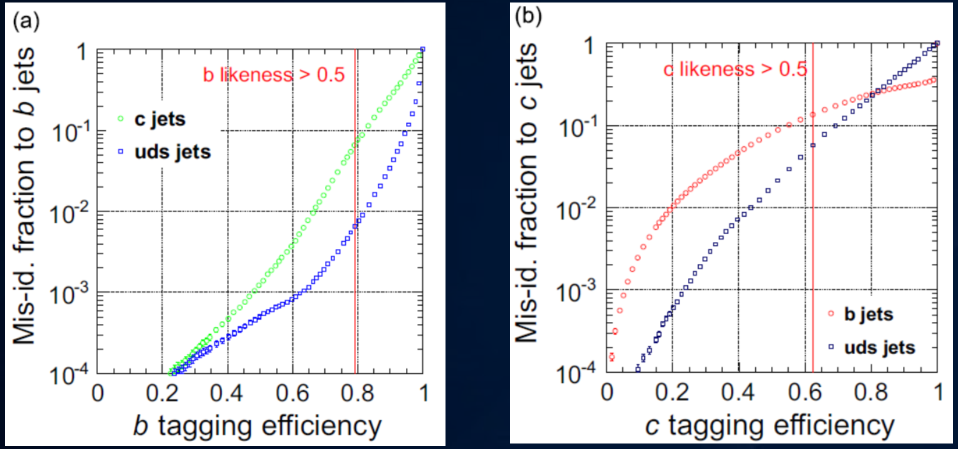
Date: 16-Aug-2023

Background

- Precise measurements instrumentation and reconstruction software are essential for the ILC PROJECT.
- Various frameworks have been developed for jet flavor identification.
- LCFIPlus (published 2013)[1] was successful in vertex finding, jet clustering and flavor tagging.
- Reached a reasonable performance of:
  - b-tag: 80% eff., 10% c / 1% uds acceptance;
  - c-tag: 50% eff., 10% b / 2% uds acceptance.

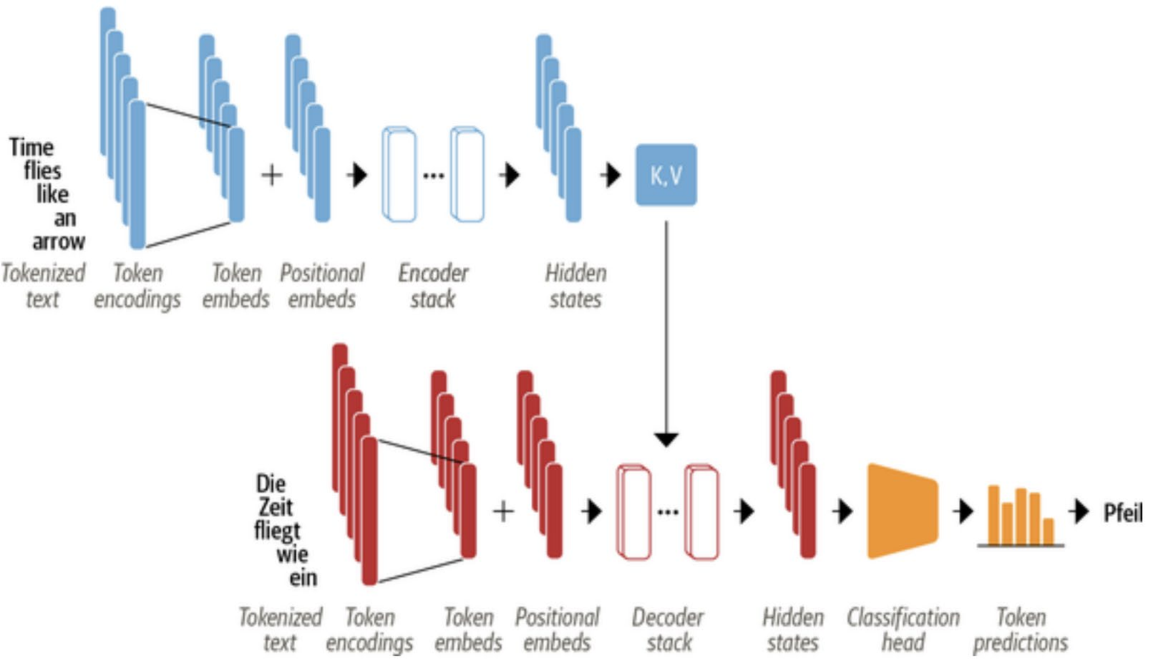


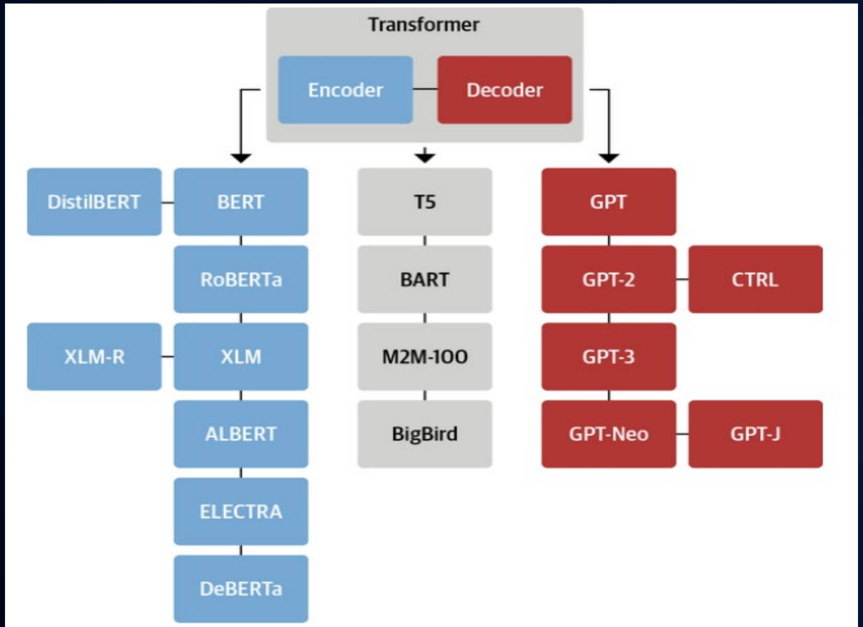
Displaced track → b/c quarks



Transformer

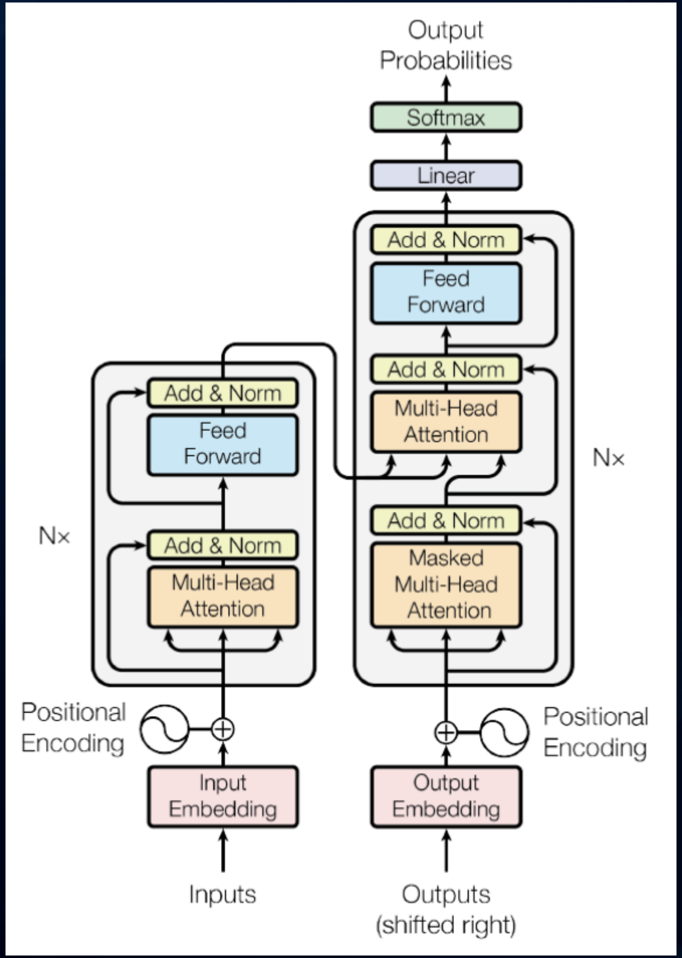
- Input is converted by the *Encoder* into a sequence of *hidden states* that is consisted of *Token Embeds* and *Positional Embeds*.
- This *hidden state* is then processed through layers of *Self-Attention* and *Feed-Forward* neural networks.
- The *Self-Attention* mechanism calculates the relative importance of each token relative to all the other tokens in the input sequence (Outperforms traditional RNN and CNN).
- The *Decoder* then outputs one token at a time, and this token is then added to the input to generate the next context iteratively.



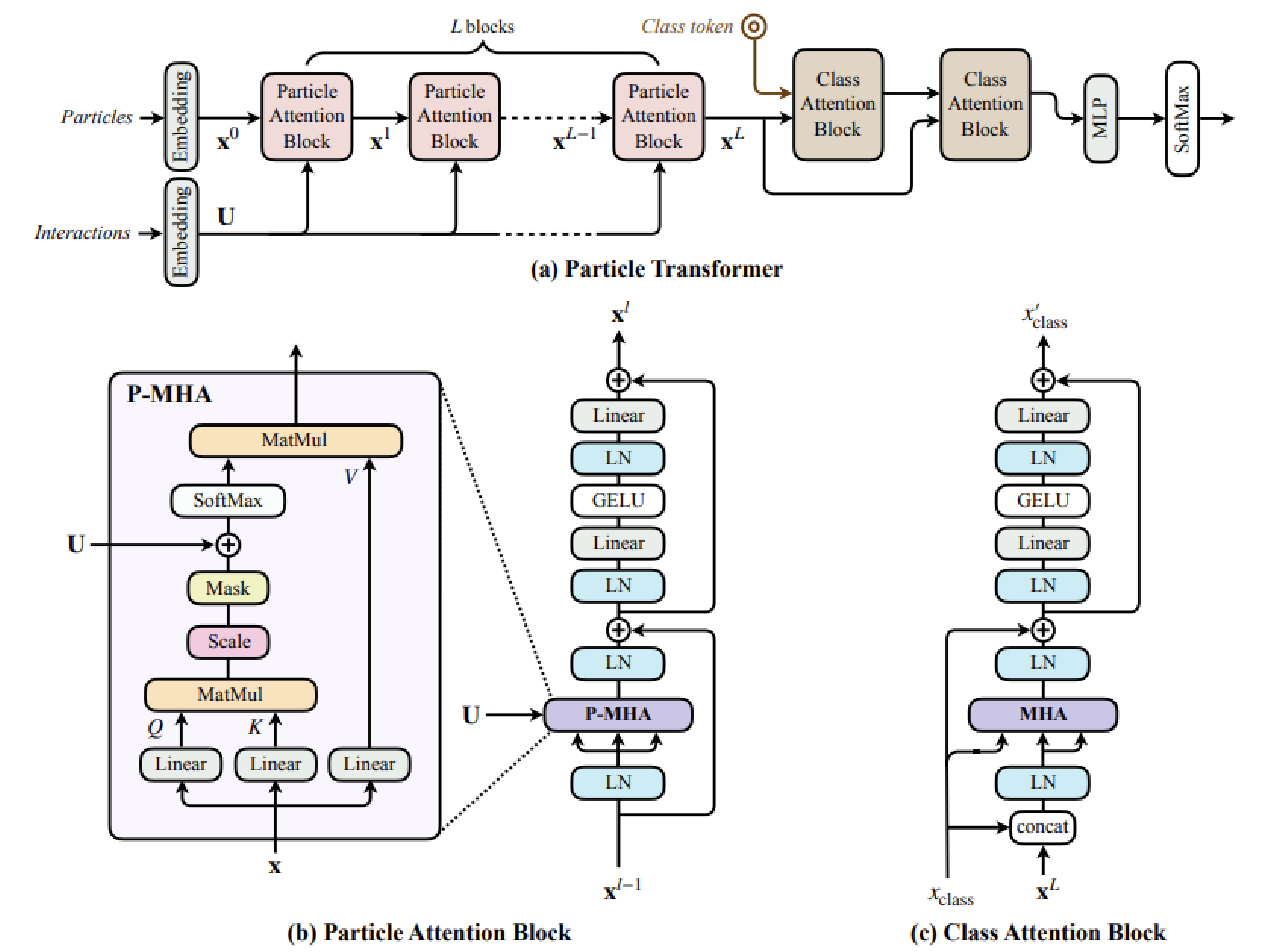


16-Aug-2023

Comparison between regular Transformer and Particle Transformer



Regular Transformer

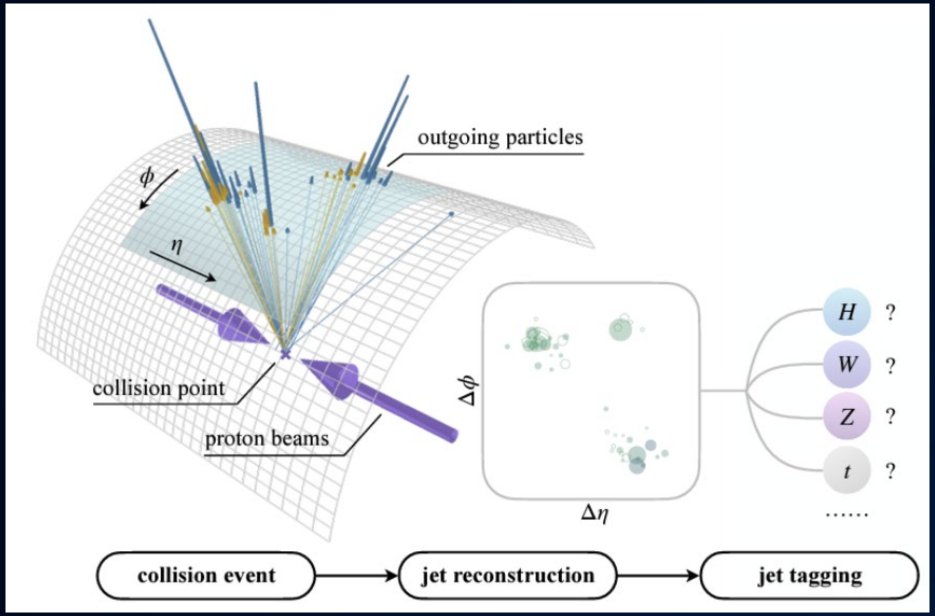


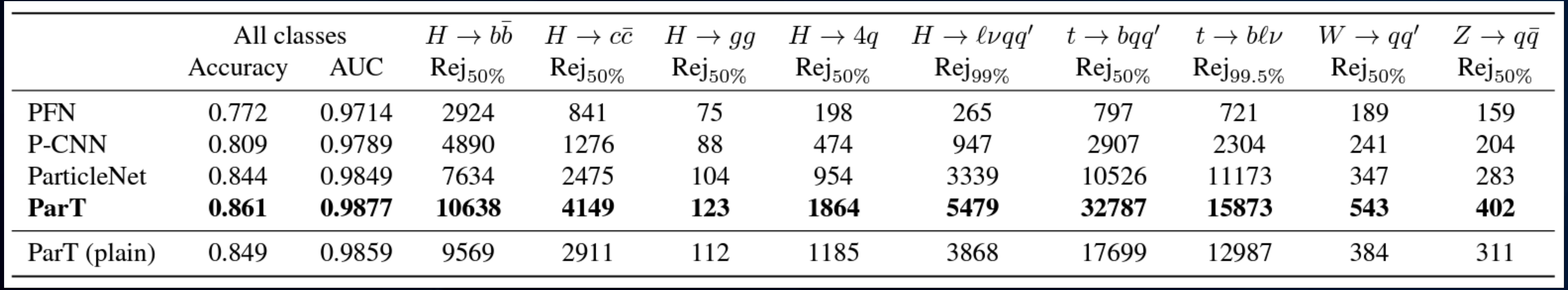
Particle Transformer

Note: MHA – MultiHeadAttention
P-MHA – Augmented version of MHA by Particle Transformer that involves Interactions Embeddings instead of Positional Embeddings

Particle Transformer (ParT)

- A new Transformer-based architecture for Jet tagging, published in 2022[2].
- It analyses the readings collected after collision events to reconstruct jets. (Illustration of CERN LHC p-p collisions)
- Surpasses the performance of previous architectures by a large margin. Values below are rejection ratio (inverse of acceptance ratio).





	All classes		$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to 4q$	$H \to \ell\nu qq'$	$t \to bqq'$	$t \to b\ell\nu$	$W \to qq'$	$Z \to q\bar{q}$
	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
PFN	0.772	0.9714	2924	841	75	198	265	797	721	189	159
P-CNN	0.809	0.9789	4890	1276	88	474	947	2907	2304	241	204
ParticleNet	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT	**0.861**	**0.9877**	**10638**	**4149**	**123**	**1864**	**5479**	**32787**	**15873**	**543**	**402**
ParT (plain)	0.849	0.9859	9569	2911	112	1185	3868	17699	12987	384	311

Data Used For Investigation

- ILD full simulation:

  1. e+ e- ⟶ qq (at 91 GeV)
     (DBD sample used for initial LCFIPlus study)
  2. e+ e- ⟶ ννH ⟶ ννqq (at 250 GeV)
     (2020 production, process ID: 410001-410006)

  With 1M jets (500k events) each

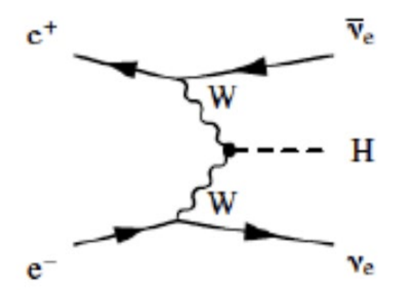
- FCCee fast simulation (Delphes with IDEA detector):

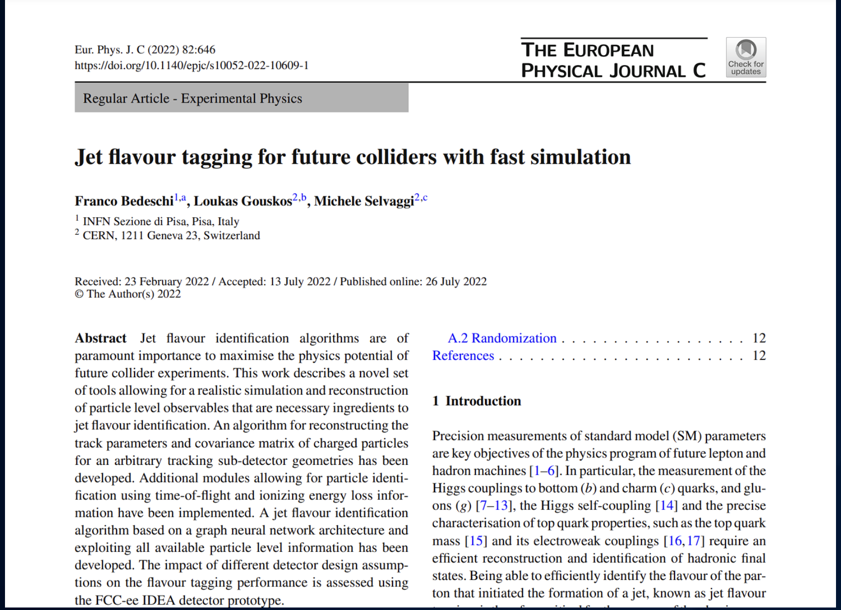
  e+ e- ⟶ ννH ⟶ ννqq (at 240 GeV)

  With 10M jets (5M events) each

- 80% are used for training, 5% for validation, 15% for test

$$\left\{ \begin{array}{c} \text{q = b,c,uds} \\ \text{ν = neutrino} \end{array} \right\}$$





Eur. Phys. J. C (2022) 82:646
https://doi.org/10.1140/epjc/s10052-022-10609-1

THE EUROPEAN PHYSICAL JOURNAL C

Check for updates

Regular Article - Experimental Physics

Jet flavour tagging for future colliders with fast simulation

Franco Bedeschi[1,a], Loukas Gouskos[2,b], Michele Selvaggi[2,c]

[1] INFN Sezione di Pisa, Pisa, Italy
[2] CERN, 1211 Geneva 23, Switzerland

Received: 23 February 2022 / Accepted: 13 July 2022 / Published online: 26 July 2022

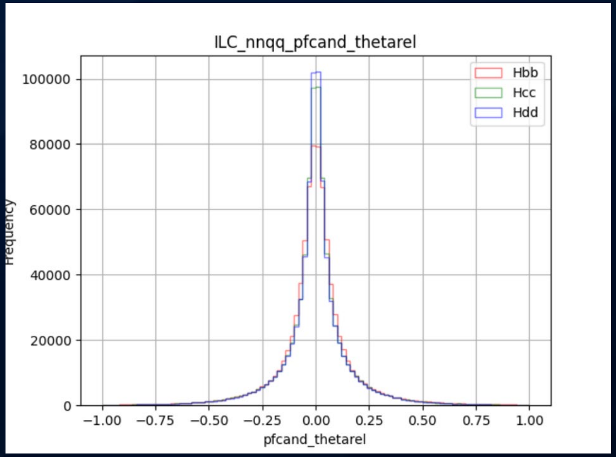
Abstract Jet flavour identification algorithms are of paramount importance to maximise the physics potential of future collider experiments. This work describes a novel set of tools allowing for a realistic simulation and reconstruction of particle level observables that are necessary ingredients to jet flavour identification. An algorithm for reconstructing the track parameters and covariance matrix of charged particles for an arbitrary tracking sub-detector geometries has been developed. Additional modules allowing for particle identification using time-of-flight and ionizing energy loss information have been implemented. A jet flavour identification algorithm based on a graph neural network architecture and exploiting all available particle level information has been developed. The impact of different detector design assumptions on the flavour tagging performance is assessed using the FCC-ee IDEA detector prototype.

1 Introduction

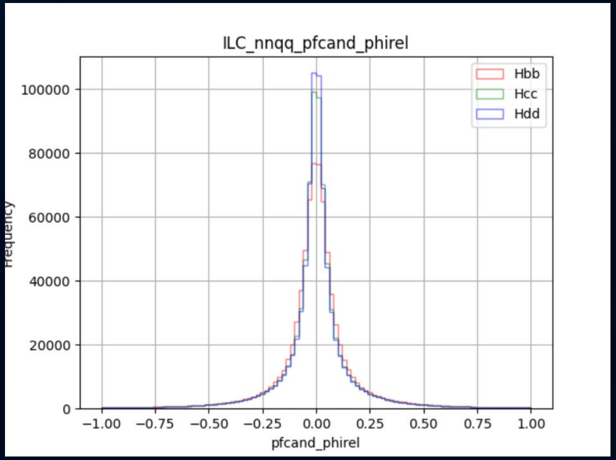
Precision measurements of standard model (SM) parameters are key objectives of the physics program of future lepton and hadron machines [1–6]. In particular, the measurement of the Higgs couplings to bottom (b) and charm (c) quarks, and gluons (g) [7–13], the Higgs self-coupling [14] and the precise characterisation of top quark properties, such as the top quark mass [15] and its electroweak couplings [16,17] require an efficient reconstruction and identification of hadronic final states. Being able to efficiently identify the flavour of the parton that initiated the formation of a jet, known as jet flavour

https://link.springer.com/article/10.1140/epjc/s10052-022-10609-1

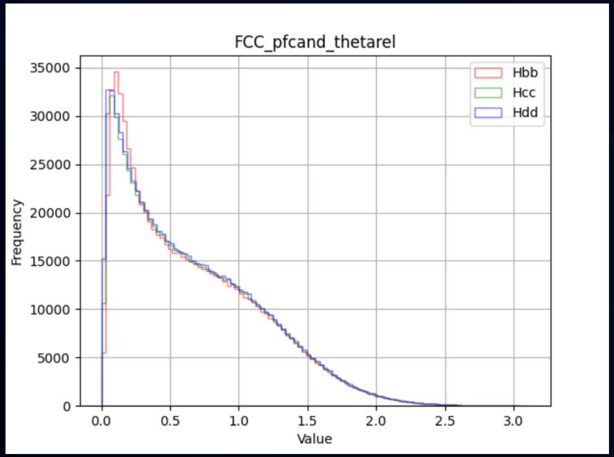
ILD vs. FCC – theta/phi distribution



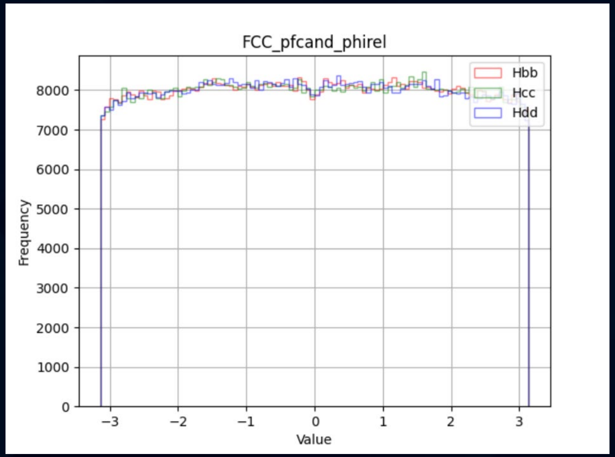
ILD theta



ILD phi



FCC theta



FCC phi

- ILD theta/phi are calculated from the difference between particle and jet theta/phi in the frame of the detector.
- FCC theta/phi are obtained from relative trace of the particle compared to the jet.
- This can cause some differences in the interaction of other parameters in the model.

Input Variables - Interactions

- FCC data uses p (scalar momentum) as interaction:

  - pfcand_p

- ILD data contains p_x, p_y, p_z (vector momentum) as interaction:

  - pfcand_px
  - pfcand_py
  - pfcand_pz

- But it's possible to transfer ILD's interaction to FCC's form for fair comparison:

$$p = \sqrt{{p_x}^2 + {p_y}^2 + {p_z}^2}$$

Input Variables - Features

- Impact Parameter (6):

pfcand_dxy
pfcand_dz
pfcand_btagSip2dVal
pfcand_btagSip2dSig
pfcand_btagSip3dVal
pfcand_btagSip3dSig

- Jet Distance (2):

pfcand_btagJetDistVal
pfcand_btagJetDistSig

- Particle ID (6):

pfcand_isMu
pfcand_isEl
pfcand_isChargedHad
pfcand_isGamma
pfcand_isNeutralHad
pfcand_type

- Kinematic (4):

pfcand_erel_log
pfcand_thetarel
pfcand_phirel
pfcand_charge

- Track Errors (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

Objectives

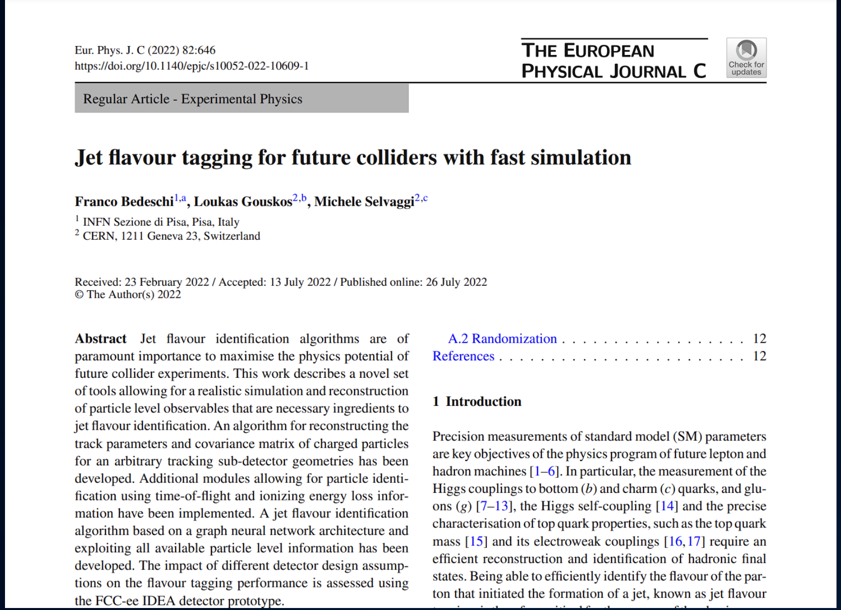
1. Confirm the performance provided with FCCee group and apply it to ILD full simulation

2. Check the performance dependence on data size and input features

3. Check origin of difference of the performance:

   - By difference on the simulation (full/fast)?

   - Detector performance?



Eur. Phys. J. C (2022) 82:646
https://doi.org/10.1140/epjc/s10052-022-10609-1

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Jet flavour tagging for future colliders with fast simulation

Franco Bedeschi[1,a], Loukas Gouskos[2,b], Michele Selvaggi[2,c]

[1] INFN Sezione di Pisa, Pisa, Italy
[2] CERN, 1211 Geneva 23, Switzerland

Received: 23 February 2022 / Accepted: 13 July 2022 / Published online: 26 July 2022

Abstract Jet flavour identification algorithms are of paramount importance to maximise the physics potential of future collider experiments. This work describes a novel set of tools allowing for a realistic simulation and reconstruction of particle level observables that are necessary ingredients to jet flavour identification. An algorithm for reconstructing the track parameters and covariance matrix of charged particles for an arbitrary tracking sub-detector geometries has been developed. Additional modules allowing for particle identification using time-of-flight and ionizing energy loss information have been implemented. A jet flavour identification algorithm based on a graph neural network architecture and exploiting all available particle level information has been developed. The impact of different detector design assumptions on the flavour tagging performance is assessed using the FCC-ee IDEA detector prototype.

1 Introduction

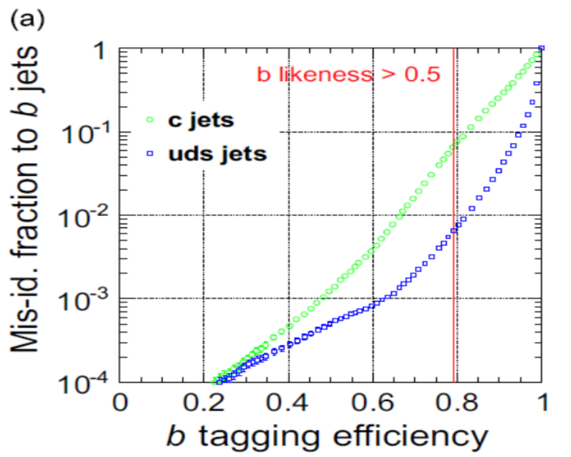
Precision measurements of standard model (SM) parameters are key objectives of the physics program of future lepton and hadron machines [1–6]. In particular, the measurement of the Higgs couplings to bottom (b) and charm (c) quarks, and gluons (g) [7–13], the Higgs self-coupling [14] and the precise characterisation of top quark properties, such as the top quark mass [15] and its electroweak couplings [16,17] require an efficient reconstruction and identification of hadronic final states. Being able to efficiently identify the flavour of the parton that initiated the formation of a jet, known as jet flavour

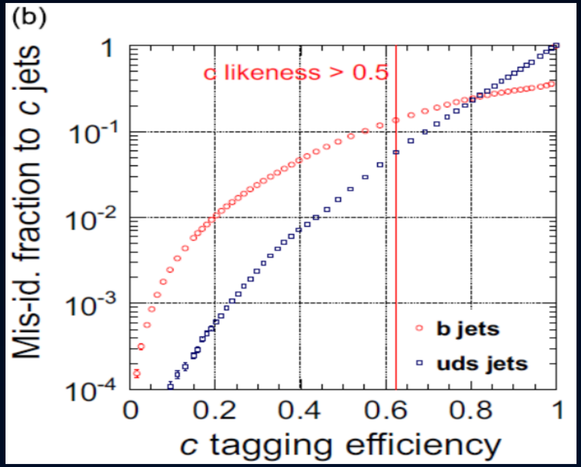
https://link.springer.com/article/10.1140/epjc/s10052-022-10609-1

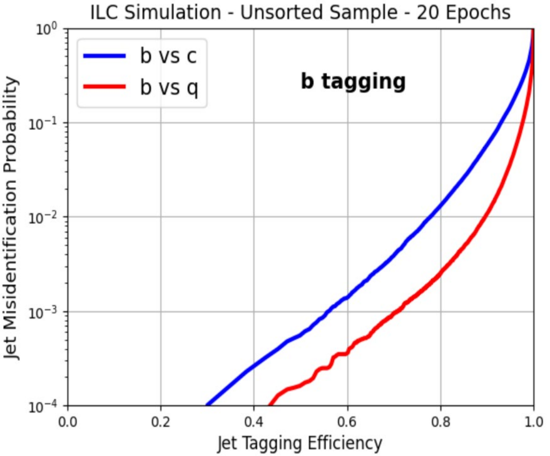
Application of ParT to ILD data

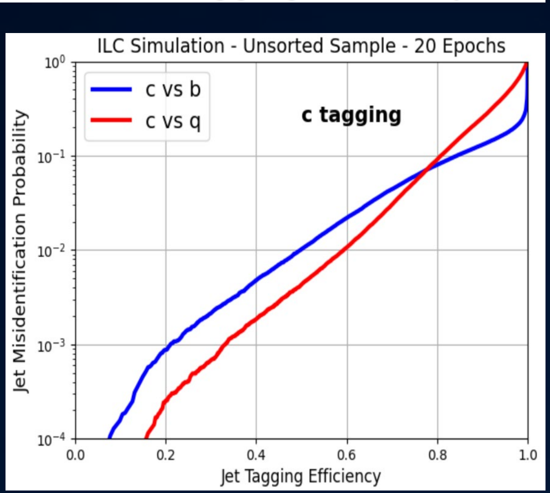
(ILD qq 91 GeV)

- Jet tagging performance is greatly improved by ParT immediately.
- The performance is improved by 4.05 – 9.80 times compared to LCFIPlus with the same set of data.
- Can this performance to be further improved?





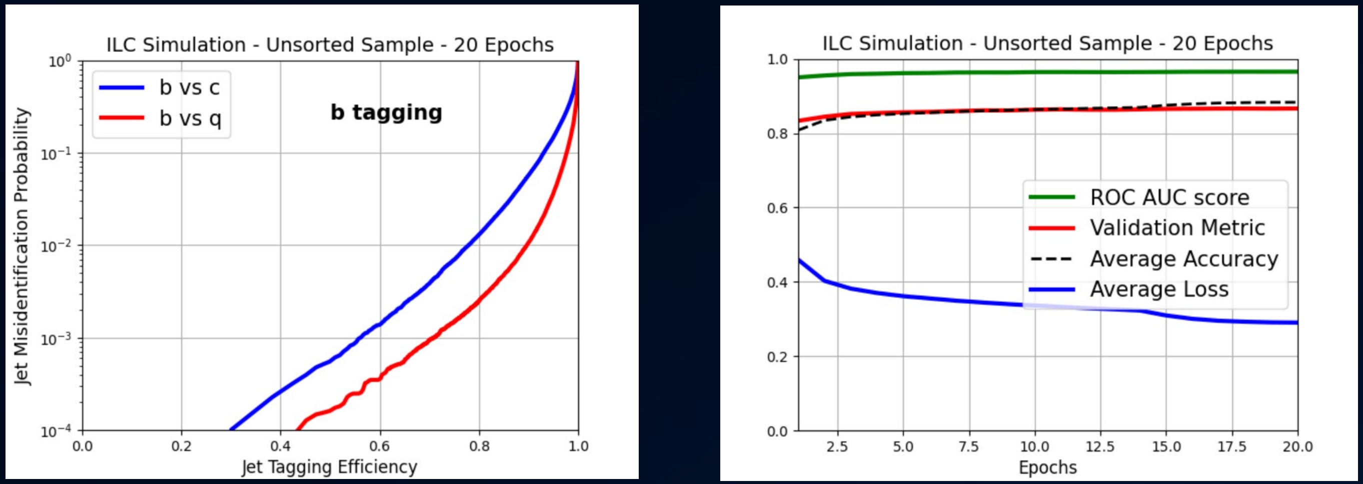




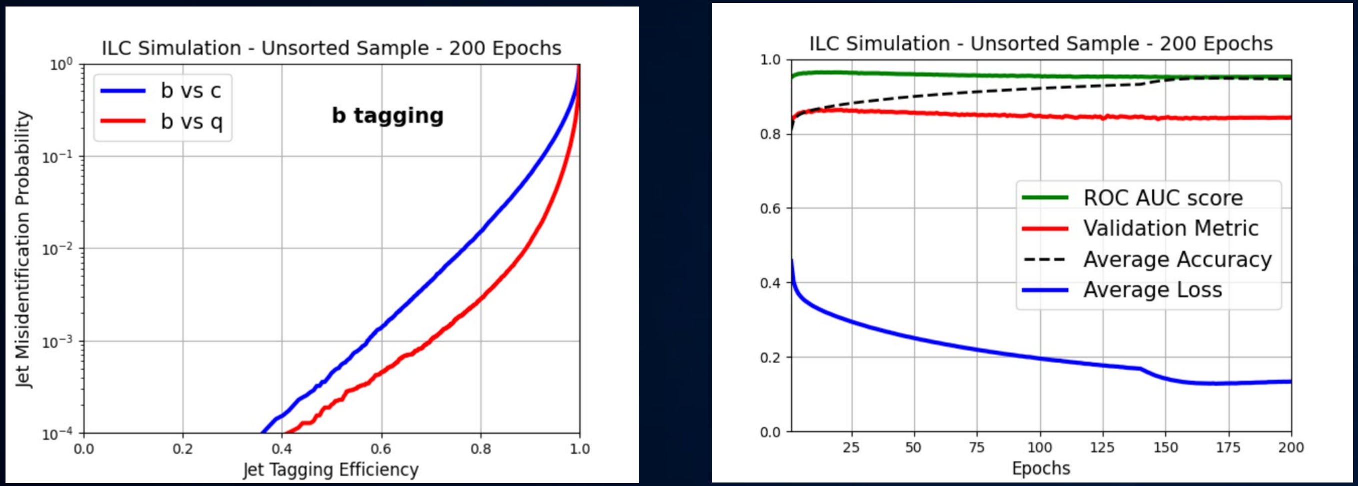
	b-tag 80% eff.		c-tag 50% eff.	
Method	c-bkg acceptance	uds-bkg acceptance	c-bkg acceptance	uds-bkg acceptance
LCFIPlus	10%	1%	10%	2%
ParT	1.29%	0.25%	1.02%	0.43%

Training parameters - epochs

- Run on NVIDIA TITAN RTX (memory: 24 GB)
  - 20 Epochs: 3 hours
  - 200 Epochs: 30 hours
- No significant improvement in tagging efficiency
- Both ROC AUC score and Validation Metric reaches a maximum around 20 epochs.
- Overtraining after 20 epochs.
- Hence 20 epochs of training is selected to avoid overtraining.



20 epochs (ILD qq 91 GeV)



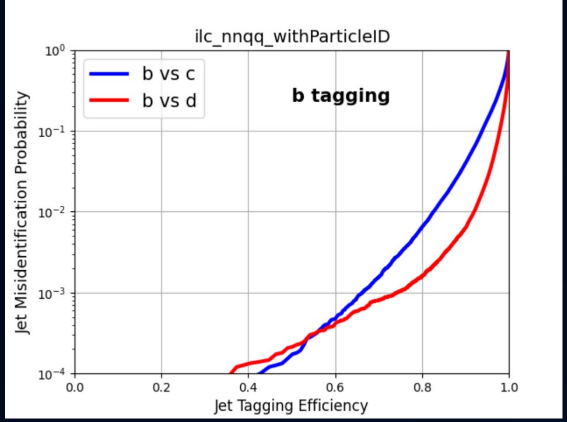
200 epochs (ILD qq 91 GeV)

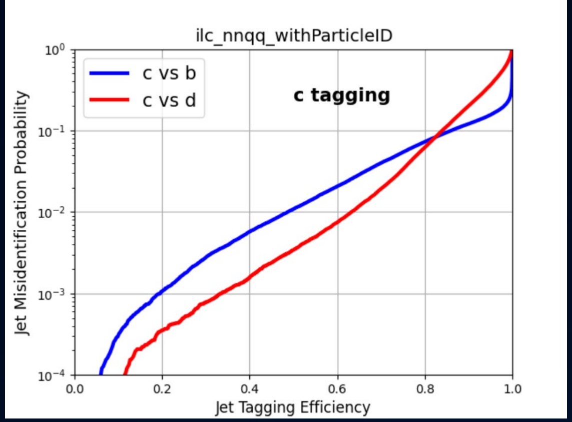
Comparison with FCC data[3]

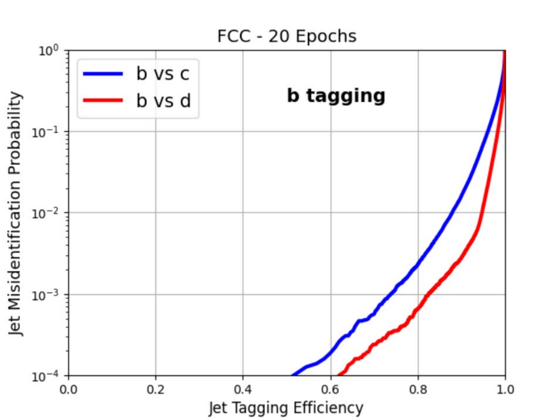
- Trained with same condition as ILD data for fair comparison. (800k data size, 20 epochs, etc.)

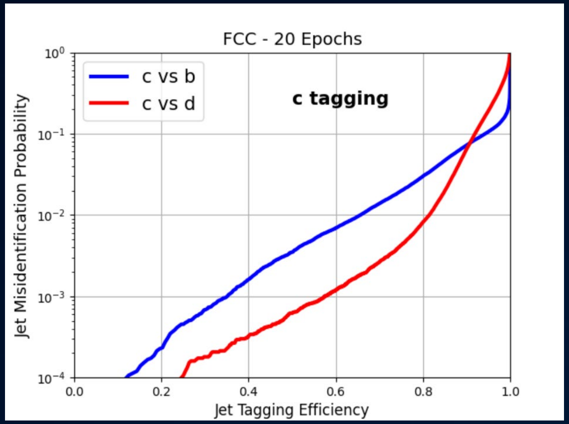
- FCC data has ~ 3 times the performance compared to ILD data.

- We would like to understand what factors caused this difference.





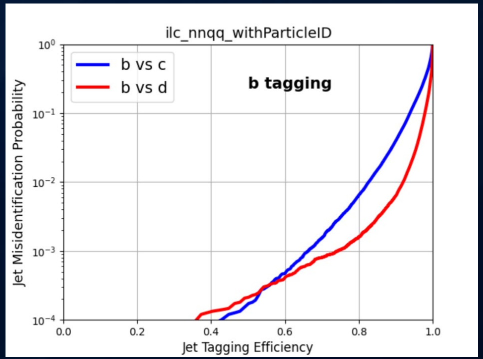




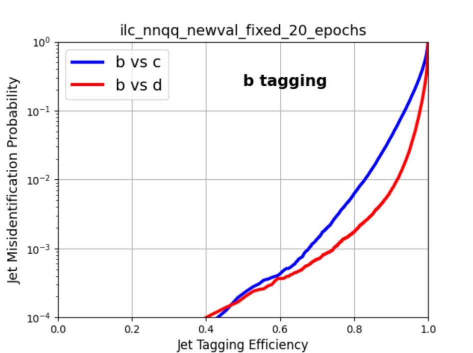
Data	Particle ID	Impact Parameters	Jet Distance	Track Errors	c-bkg acceptance @ b-tag 80% eff.	b-bkg acceptance @ c-tag 50% eff.
ILD **(vvqq 250 GeV)**	◍	◍	◍	◍	**0.64%**	**1.09%**
FCC	◍	◍	◍	◍	**0.23%**	**0.35%**

Effect of different parameters: ILD (vvqq 250 GeV)

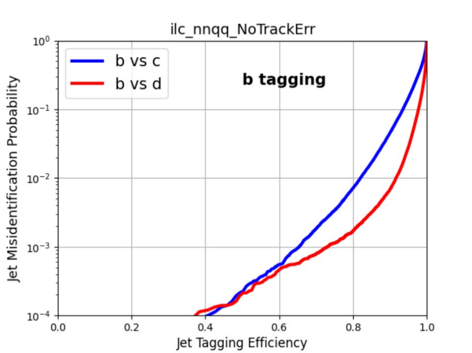
(1)



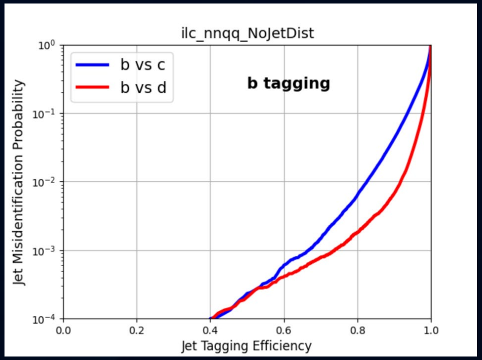
(2)



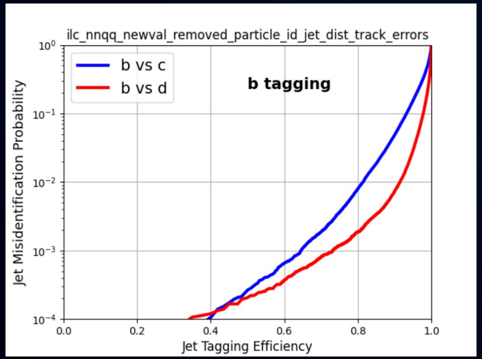
(3)



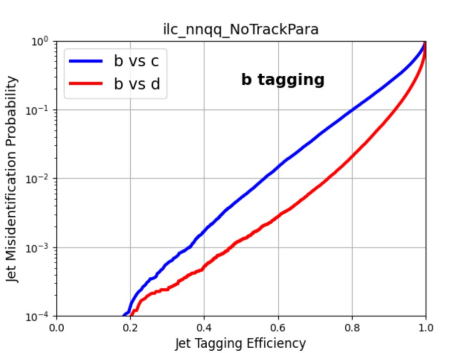
(4)



(5)



(6)

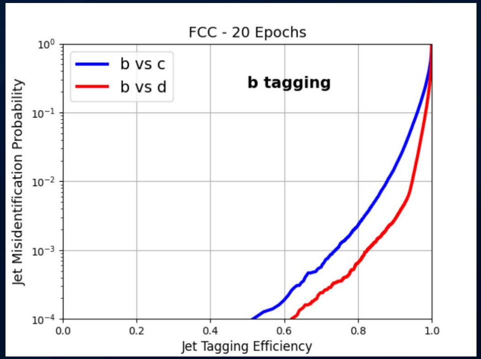


Plot Index	Particle ID	Impact Parameters	Jet Distance	Track Errors	c-bkg acceptance @ b-tag 80% eff.	b-bkg acceptance @ c-tag 50% eff.
(1)	◍	◍	◍	◍	0.64%	1.09%
(2)	✗	◍	◍	◍	0.62%	1.14%
(3)	✗	◍	◍	✗	0.71%	1.24%
(4)	✗	◍	✗	◍	0.63%	1.19%
(5)	✗	◍	✗	✗	0.79%	1.28%
(6)	✗	✗	◍	◍	9.69%	6.91%

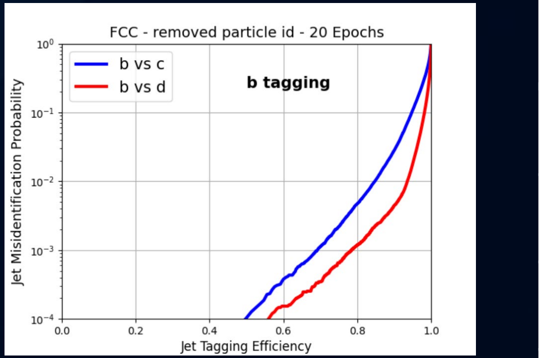
- Impact parameter gives most significance in affecting the training performance.
- The other parameters are about the similar significance (not significant impact).

Effect of different parameters: FCC

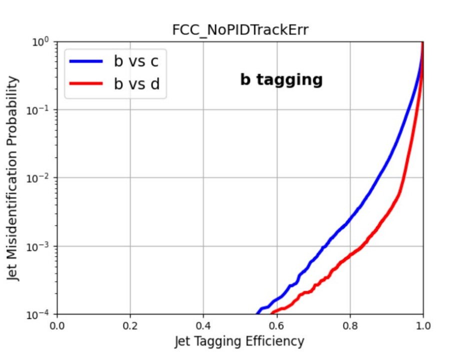
(1)



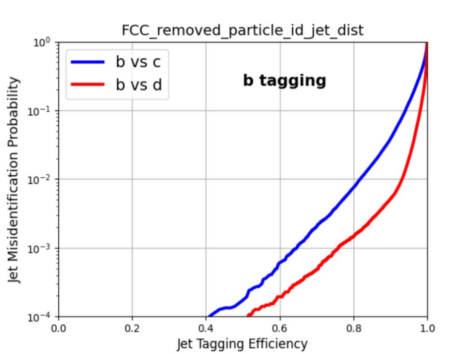
(2)



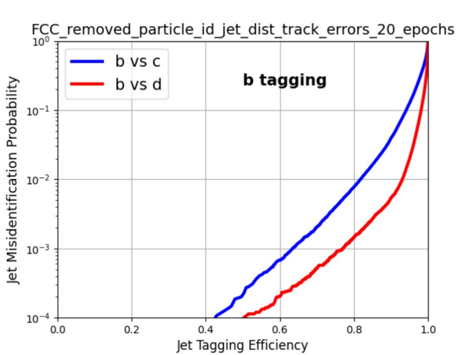
(3)



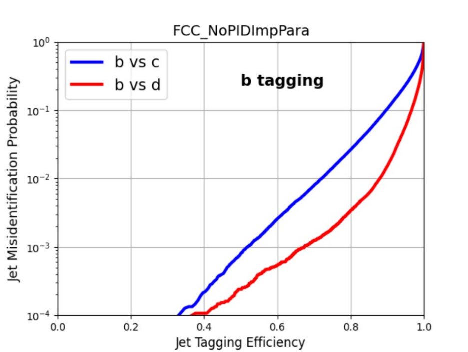
(4)



(5)



(6)



Plot Index	Particle ID	Impact Parameters	Jet Distance	Track Errors	c-bkg acceptance @ b-tag 80% eff.	b-bkg acceptance @ c-tag 50% eff.
(1)	◍	◍	◍	◍	0.23%	0.35%
(2)	✗	◍	◍	◍	0.47%	0.64%
(3)	✗	◍	◍	✗	0.24%	0.35%
(4)	✗	◍	✗	◍	0.75%	0.80%
(5)	✗	◍	✗	✗	0.77%	0.80%
(6)	✗	✗	◍	◍	2.64%	1.58%

- Effect of Impact Parameters also significant.

- Both Particle ID and Jet Distance give significant impacts.

- Removal of track errors improves performance, could be a result of too many variables of Track Errors (15) shifting away the contribution of others. Further investigation should be conducted.

ILD (vvqq 250 GeV) vs. FCC

					c-bkg acceptance @ b-tag 80% eff.		b-bkg acceptance @ c-tag 50% eff.	
Plot Index	Particle ID	Impact Parameters	Jet Distance	Track Errors	ILD	FCC	ILD	FCC
(1)	◍	◍	◍	◍	0.64%	0.23%	1.09%	0.35%
(2)	✕	◍	◍	◍	0.62%	0.47%	1.14%	0.64%
(3)	✕	◍	◍	✕	0.71%	0.24%	1.24%	0.35%
(4)	✕	◍	✕	◍	0.63%	0.75%	1.19%	0.80%
(5)	✕	◍	✕	✕	0.79%	0.77%	1.28%	0.80%
(6)	✕	✕	◍	◍	9.69%	2.64%	6.91%	1.58%

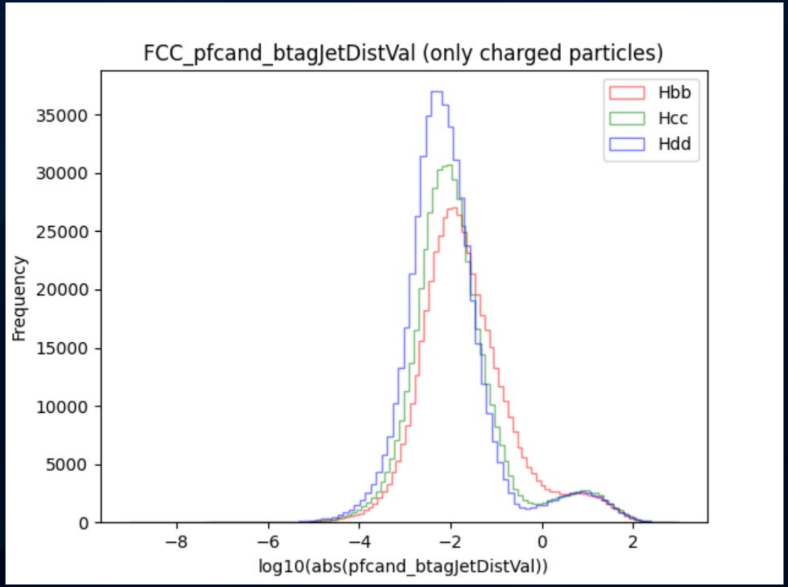
- Overall, ILD data is performing slightly worse than FCC data in ParT training.

- This could be:
  1. FCC has rather ideal detector response as a result of fast simulation
  2. FCC’s Impact Parameter has potentially better resolution
  3. The Particle ID of ILD is rather simple, not yet including the recent development

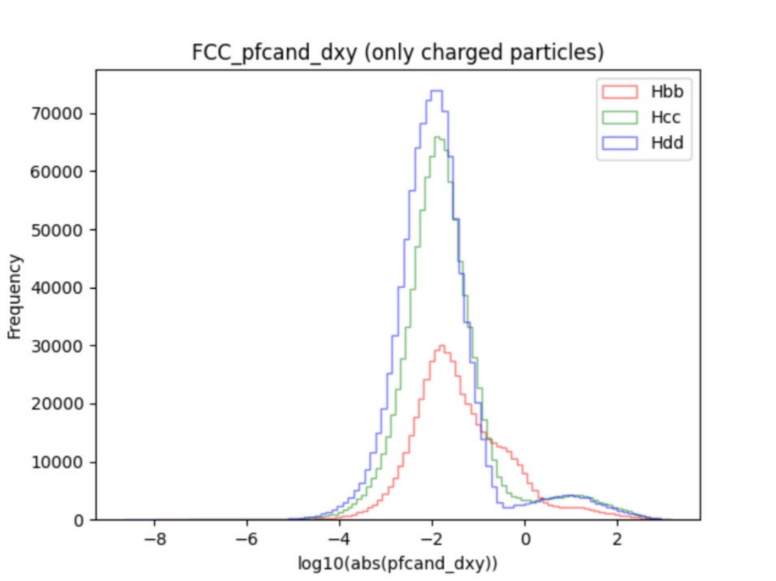
- For (5), when the input variable is reduced to be only Impact Parameters, the performance for b-tagging becomes very similar, while FCC does better in c-tagging

- This potentially indicates that resolution of Impact Parameter is more crucial for c-tagging than b-tagging (since charm hadrons decay faster than heavier bottom hadrons)

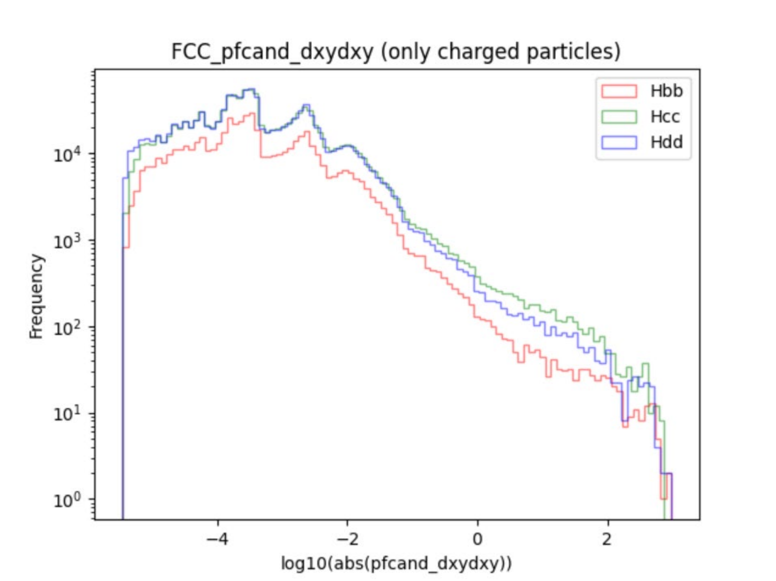
Potential Improvement: log(abs)



Jet Distance



Impact Parameter



Track Errors

- Some example distribution of log(abs) the three parameters
- All very small (largely gathering around 10^{-2})
- Hence log(abs) potentially spreads out the distribution and make it more readable by the architecture
- Can potentially improve the performance?

Potential Improvement: log(abs)

Particle ID	Impact Parameters	Jet Distance	Track Errors	c-bkg acceptance @ b-tag 80% eff.	b-bkg acceptance @ c-tag 50% eff.
✕	◍	◍	◍	**0.62%**	**1.14%**
✕	◍ +log(abs)	◍ +log(abs)	◍ +log(abs)	**0.54%**	**1.06%**
✕	◍	◍ +log(abs)	◍ +log(abs)	**0.79%**	**1.33%**
✕	◍	◍ +log(abs)	◍	**0.78%**	**1.36%**
✕	◍ +log(abs)	◍	◍	**0.47%**	**1.03%**
✕	log(abs)	log(abs)	log(abs)	**0.82%**	**1.32%**
✕	◍	log(abs)	log(abs)	**0.80%**	**1.37%**
✕	◍	◍	log(abs)	**0.82%**	**1.38%**

- Adding log(abs) to three parameters of ILD (vvqq 250 GeV) does improve performance.
- However, the addition of log(abs) of Jet Distance and Track Errors only decreases the performance.
- Can be a result of too many parameters lowers the weight of contribution of impact parameter in the model, which is more significant.
- Addition of only log(abs) of Impact Parameters gives the best performance.
- Also tried replacing the original values with log(abs).
- Performance decreased – possible loss of directional information.

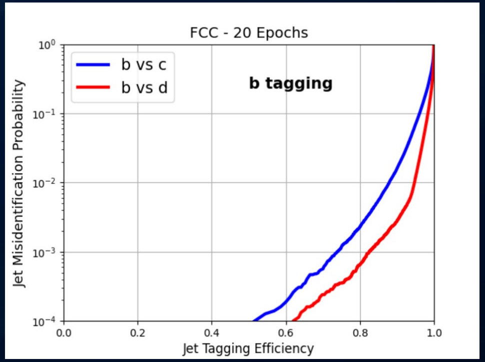
Use px, py, pz instead of p (Interaction)

				c-bkg acceptance @ b-tag 80% eff.		b-bkg acceptance @ c-tag 50% eff.	
Particle ID	Impact Parameters	Jet Distance	Track Errors	p	p_x p_y p_z	p	p_x p_y p_z
✕	◍	◍	◍	0.62%	0.49%	1.14%	1.01%
✕	◍ +log(abs)	◍ +log(abs)	◍ +log(abs)	0.54%	0.52%	1.06%	1.00%
✕	◍ +log(abs)	◍	◍	0.47%	0.50%	1.03%	0.97%

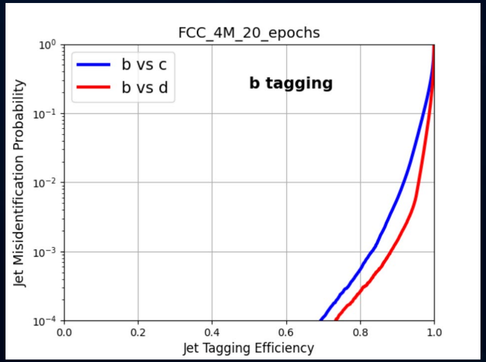
- ILD (vvqq 250 GeV) data shows that application of px, py, pz has better performance than p.
- However, application of log(abs) of the parameters becomes less significant.
- Can be because that application of px, py, pz changes the way log(abs) interacts with other parameters.
- Other potential treatments can be investigated.

Sample size affects performance

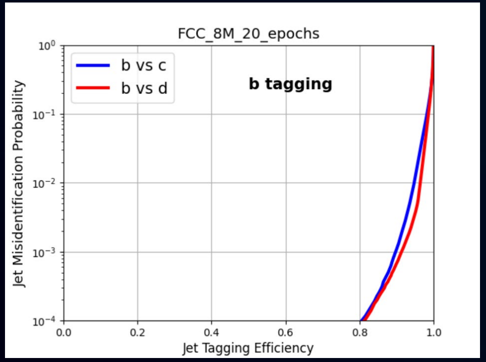
(1)



(2)



(3)



Plot Index	Particle ID	Impact Parameters	Jet Distance	Track Errors	Training Sample size	c-bkg acceptance @ b-tag 80% eff.	b-bkg acceptance @ c-tag 50% eff.
(1)	◍	◍	◍	◍	800k	0.23%	0.35%
(2)	◍	◍	◍	◍	4M	0.054%	0.20%
(3)	◍	◍	◍	◍	8M	0.0076%	0.10%

- Training performance significantly improved with bigger data sample size
- Training sample size change of FCC data:

  800k -> 4M: 4 times better performance (b-tagging)

  4M -> 8M: 5 times better performance (b-tagging)
- This non-linearity of increase in performance should be further investigated.
- Bigger data size of ILD should be obtained for better performance, as well as comparison with FCC data for further investigation on its behaviour.

Training parameters – More to be confirmed

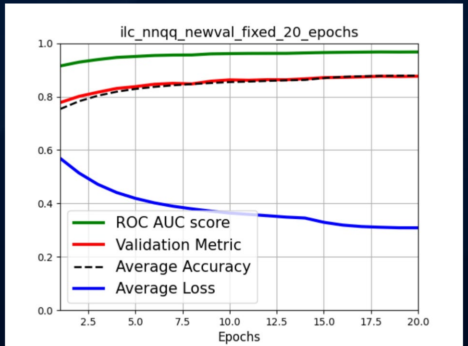
Fine tuning

							c-bkg acceptance @ b-tag 80% eff.		b-bkg acceptance @ c-tag 50% eff.	
Particle ID	Impact Parameters	Jet Distance	Track Errors	Fine-Tuning Sample	Training Sample	Similar theta/phi ?	No Fine-Tuning	With Fine-Tuning	No Fine-Tuning	With Fine-Tuning
✕	◍	◍	◍	FCC 240 GeV (8M)	ILD 250 GeV (800k)	◍	0.62%	1.37%	1.14%	1.95%
✕	◍	◍	◍	FCC 240 GeV (8M)	ILD 250 GeV (800k)	✕	1.77%	1.32%	2.22%	2.01%
◍	◍	◍	◍	ILD 250 GeV (800k)	ILD 91 GeV (80k)	◍	4.49%	0.97%	3.79%	1.53%

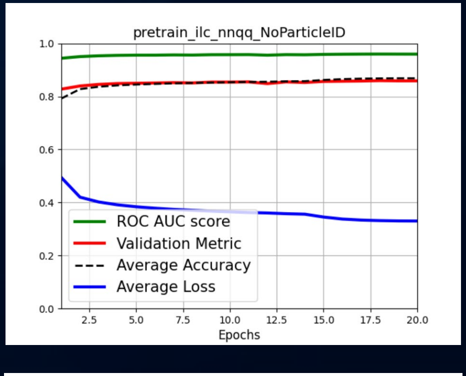
- Use result of 8M FCC data to train ILD 800k data
- Improves performance only when setups are similar
- Training of same setup (pretrain ILD 91 GeV data with ILD 250 GeV data) gives best performance
- Further investigation should be conducted on how to maximise the outcome for fine-tuning between different data sets

Fine tuning – Training curves

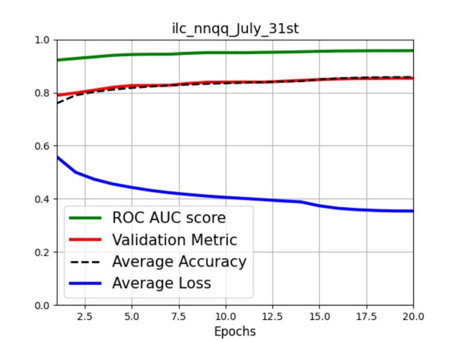
(1)



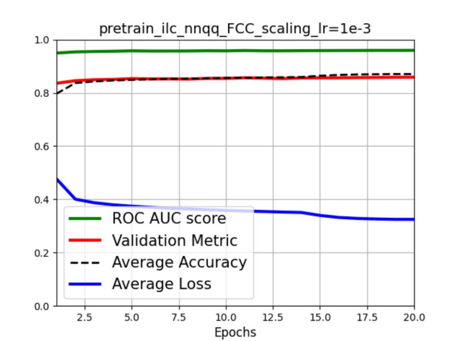
(2)



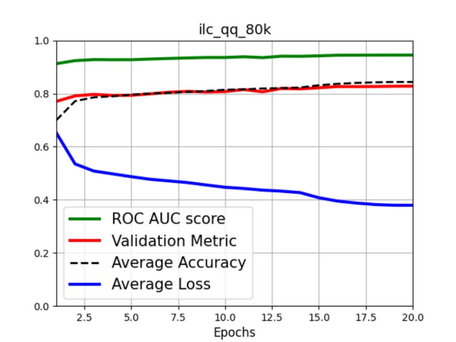
(3)



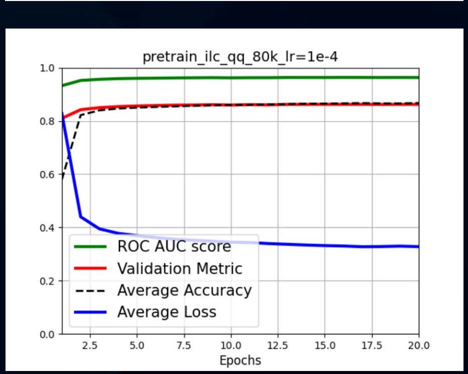
(4)



(5)



(6)



							Plot Indices	
Particle ID	Impact Parameters	Jet Distance	Track Errors	Fine-Tuning Sample	Training Sample	Similar theta/phi?	No Fine-Tuning	With Fine-Tuning
✕	▨	▨	▨	FCC 240 GeV (8M)	ILD 250 GeV (800k)	▨	(1)	(2)
✕	▨	▨	▨	FCC 240 GeV (8M)	ILD 250 GeV (800k)	✕	(3)	(4)
▨	▨	▨	▨	ILD 250 GeV (800k)	ILD 91 GeV (80k)	▨	(5)	(6)

- With fine-tuning, the training is obviously accelerated for the initial epochs (even for those with worse eventual performance)

- This is particularly obvious for plot (6) – similar simulation setup data

Potential Further Investigation

1. Application to real physics data (e.g. Higgs identification)
2. Potentially combine LCFIPlus with ParT to further improve performance
3. Train with bigger sample of ILD
4. Fast simulation data of ILD can be potentially used for pretraining for the full simulation data
5. Particle ID for ILD data can be better implemented by applying the timing and dE/dx measurement (can also be used for testing accuracy of detectors required by examining the strange-tagging performance)
6. Applying transformer to other reconstruction algorithms (e.g. particle flow) and investigate on its wider usage

Summary

- Particle Transformer is a very promising in quark flavour tagging.
- Its performance can be further improved by adjusting the input parameters.
- Bigger data set is required for better training outcomes.
- Fine-tuning is effective with the model, but only for similar data setups.
- Its application on other reconstruction algorithms should be explored.

Reference List

[1] https://doi.org/10.1016/j.nima.2015.11.054

[2] https://arxiv.org/abs/2202.03772

[3] https://link.springer.com/article/10.1140/epjc/s10052-022-10609-1