Strange tagging with ILD full simulation and application to H->ss analysis

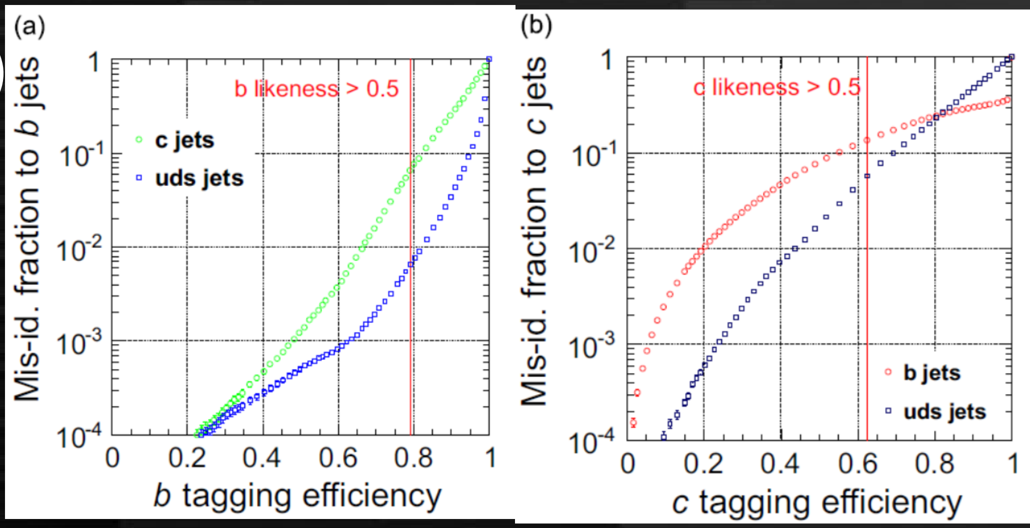
Taikan Suehara / 末原　大幹
(ICEPP, The University of Tokyo)

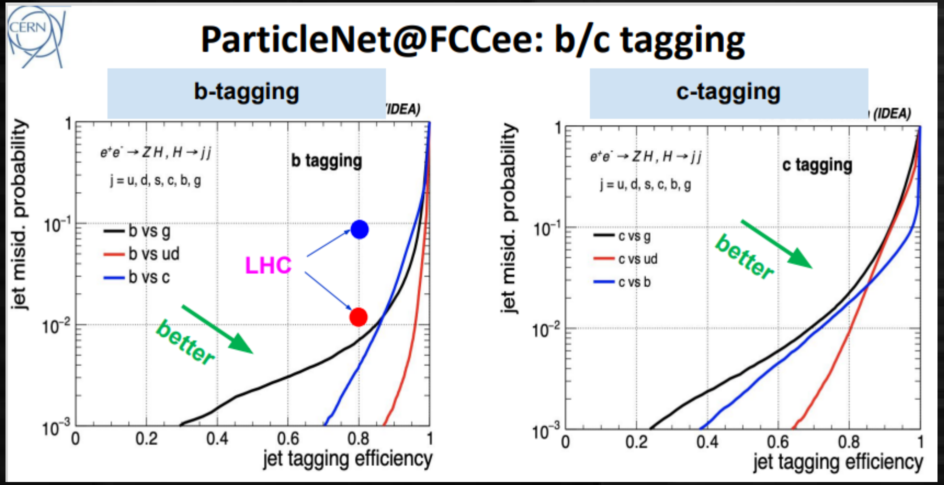
R. Tagami (ICEPP, Tokyo), T. Tanabe (MI-6 Ltd.),
R. Sugawara, R. Hosokawa, S. Narita (Iwate U.),

Flavor tagging for Higgs factories

- Jet flavor tagging is essentially important for Higgs studies (including self coupling)
- LCFIPlus (published 2013) was long used for flavor tagging
  - All physics performance in ILD/SiD/CLIC are based on LCFIPlus
- FCCee reported >10x better rejection using ParticleNet (GNN) in 2022
  - Delphes is used for simulation
- We studied DNN-based flavor tag with ILD full simulation to confirm it
  - Using latest algorithm: Particle Transformer (ParT)

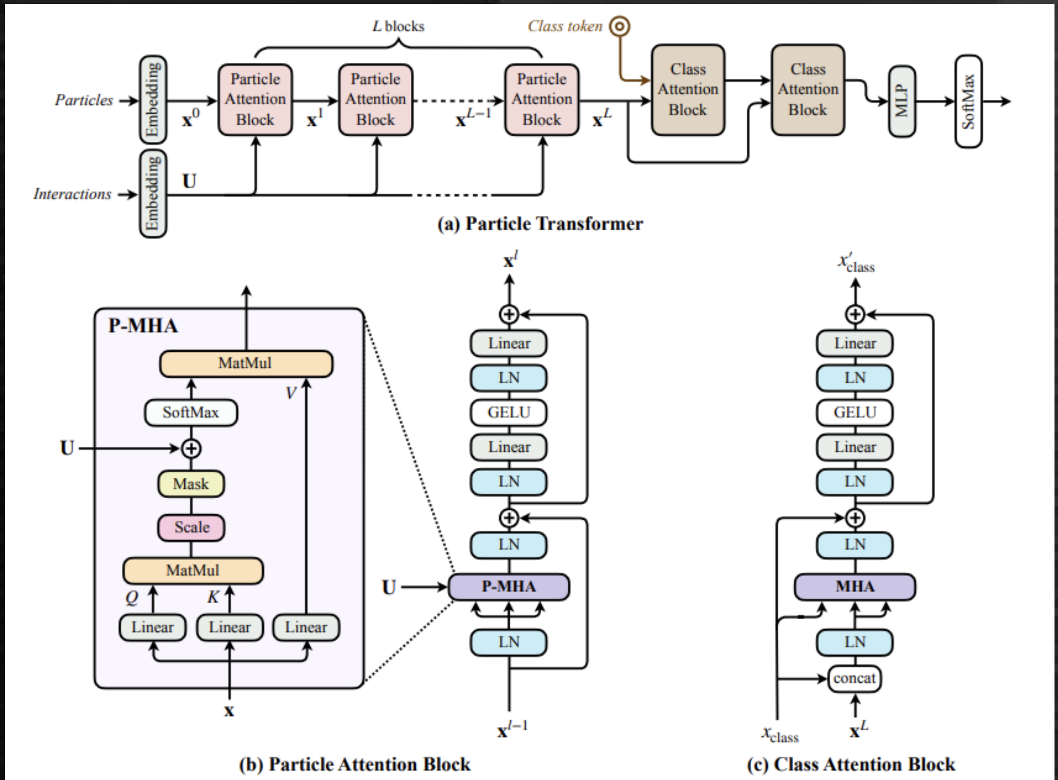
LCFIPlus performance plots

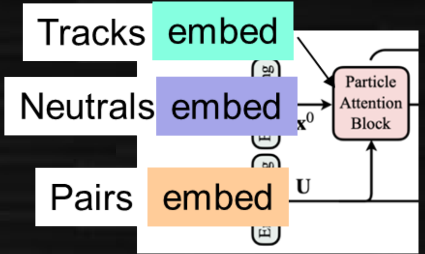




Particle Transformer (ParT)

- Transformer: self-attention-based algorithm intensively used for NLP (e.g. chatGPT)
  - Weak biasing: possible to train big samples efficiently (with more learnable weights) but demanding big training sample for high performance
- ParT is a new Transformer-based architecture for Jet tagging, published in 2022.
  - Pair-wise variable (angle, mass etc.) is added to plain Transformer encoder to boost attention
- Surpasses the performance of ParticleNet
  - ParticleNet only looks "neighbor" particles while Transformer uses attention to learn where to look





	All classes	
	Accuracy	AUC
PFN	0.772	0.9714
P-CNN	0.809	0.9789
ParticleNet	0.844	0.9849
ParT	**0.861**	**0.9877**

Performance with JetClass event classification (100M sample)

Data Samples and Input Variables

Data samples

- ILD full simulation

  1. e+ e- → qq (at 91 GeV) (used in LCFIPlus study)
  2. e+ e- → ννH → ννjj (at 250 GeV) (2020 production)

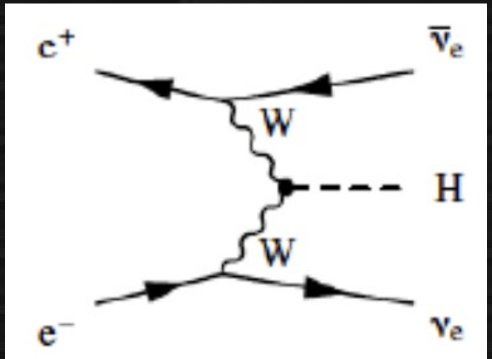
q = b,c,uds
j=b,c,u,d,s,g

1M jets (500k events) for each flavor

- FCCee fast simulation (Delphes with IDEA detector):

  e+ e- → ννH → ννjj (at 240 GeV)

10M jets (5M events) each flavor



80% for training
5% for validation
15% for test

Input variables

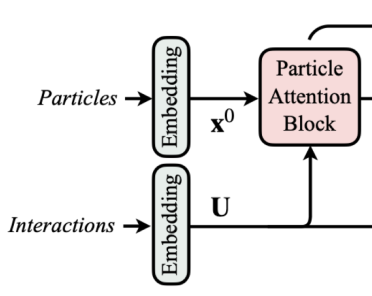
Particles: for every track/neutral

- Impact parameters (6)
  - 2D/3D, from primary vertex
- Jet distance (2)
  - Displacement from jet axis
- Covariant matrix (15)
- Kinematics (4)
  - Energy fraction, angles, charge
- Particle ID (6)
  - Probability (or binary selection) of e, μ, hadron, gamma, neutral hadron

Interactions: for every particle pair

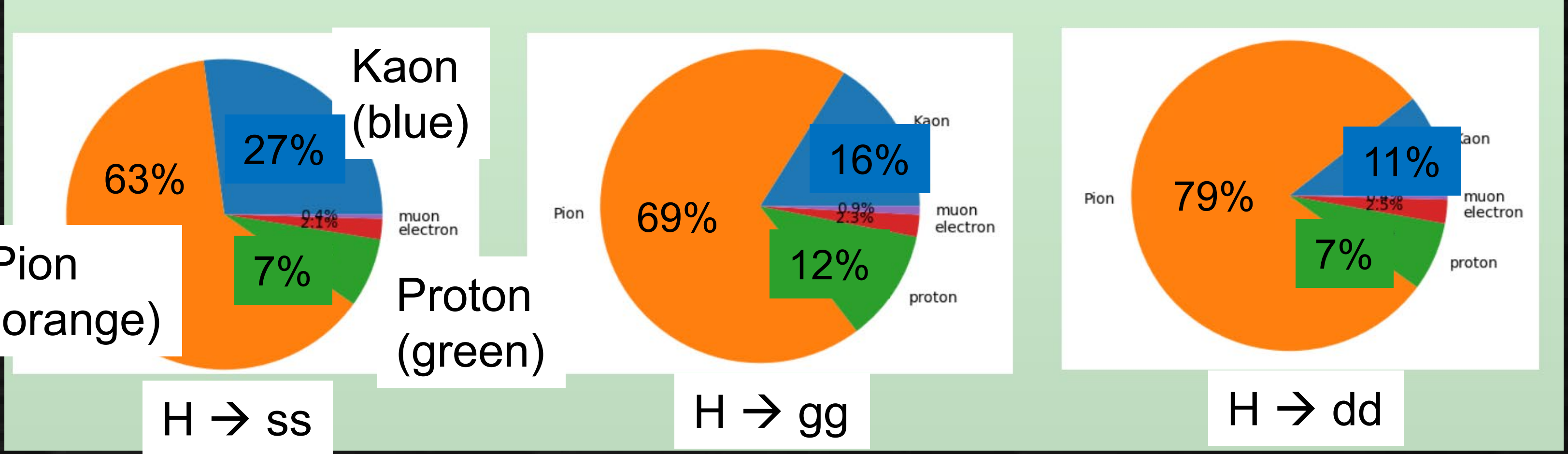
- δR^2, k_t, Z, mass

Input of ParT



Strange tagging

- High-momentum kaon in jet is a clue to strange jets
  - Contamination from g→ss give relatively low momentum
- dE/dx is essential for Particle ID in ILD
  - As well as ToF, but only effective in low energy tracks (which are less important in strange tagging)
- Using newly-developed comprehensive PID
  - Giving much better separation than previous PID



Fractions of tracks having > 5 GeV

More Kaons in ss
More protons in gg

Fraction of true particles

True particle (columns) / CPID prediction (rows)

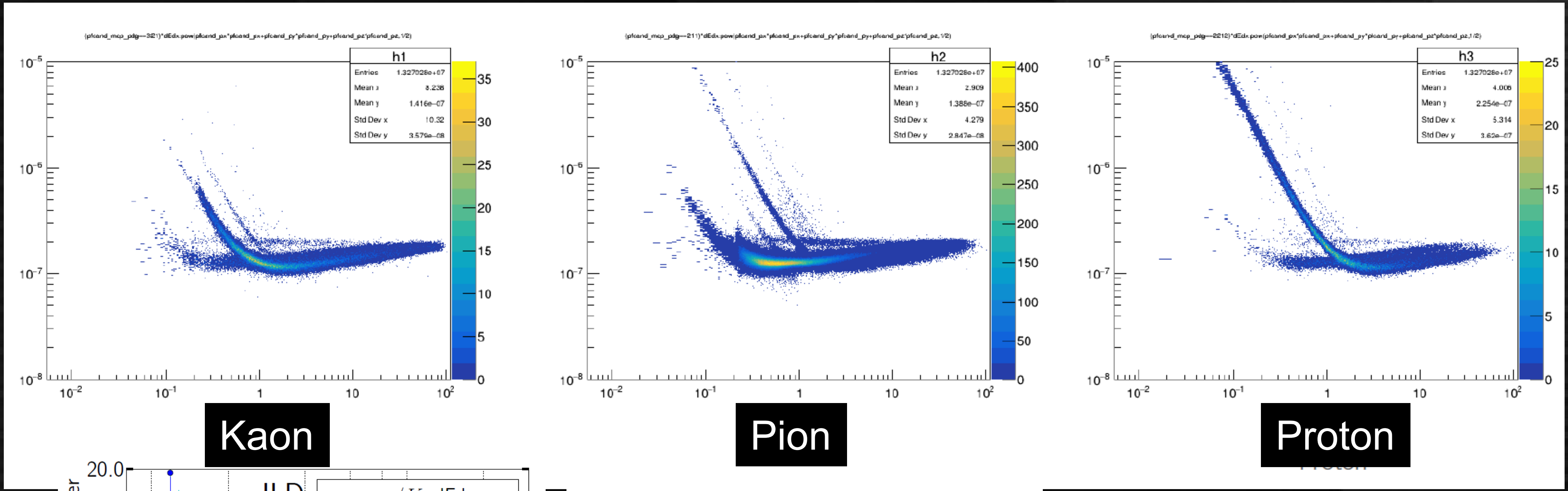
↑ 3 < p < 5 GeV

	K	π	p	e	μ
K	0.65	0.04	0.20	0.04	0.10
π	0.08	0.90	0.04	0.32	0.28
p	0.26	0.04	0.76	0.09	0.08
e	0.00	0.00	0.00	0.53	0.01
μ	0.01	0.02	0.00	0.01	0.53

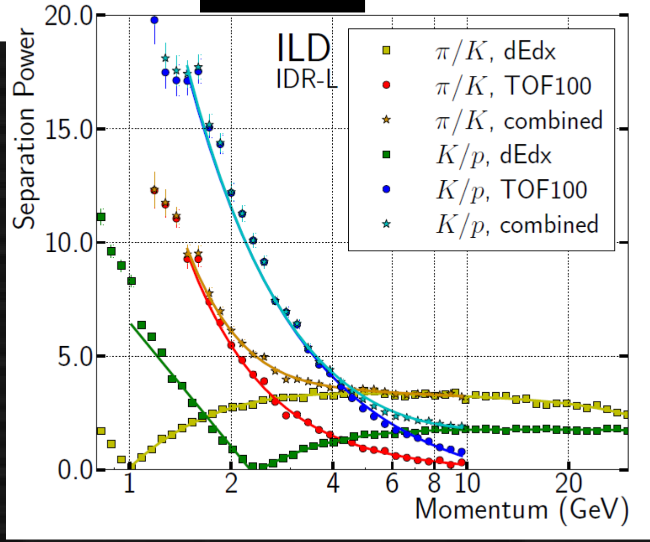
↑ p > 5 GeV

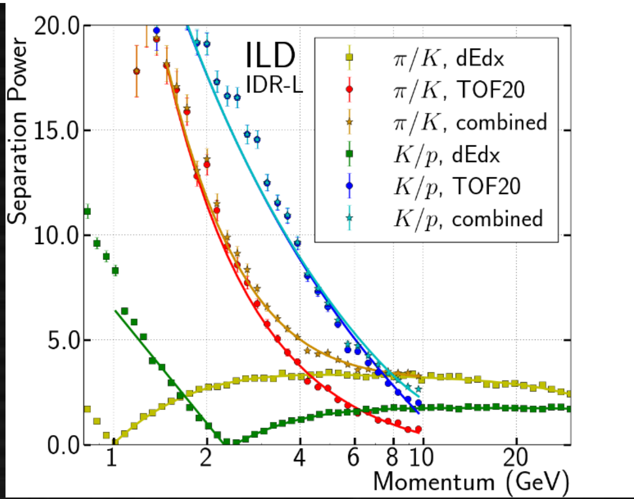
	K	π	p	e	μ
K	0.74	0.07	0.20	0.13	0.16
π	0.07	0.89	0.03	0.40	0.37
p	0.18	0.03	0.76	0.09	0.06
e	0.00	0.00	0.00	0.38	0.01
μ	0.01	0.01	0.00	0.01	0.40

dE/dx distribution and separation power



dE/dx inside strange jets (separated by MC PID)





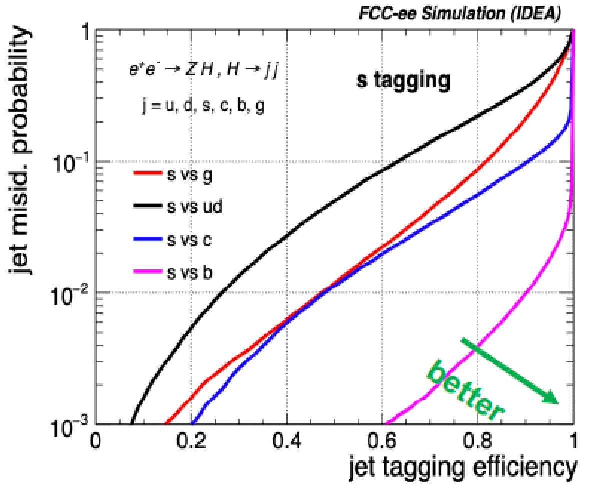
PID by ILD
TOF and dEdx combined
20/100 psec resolution assumed

Strange tagging: initial results

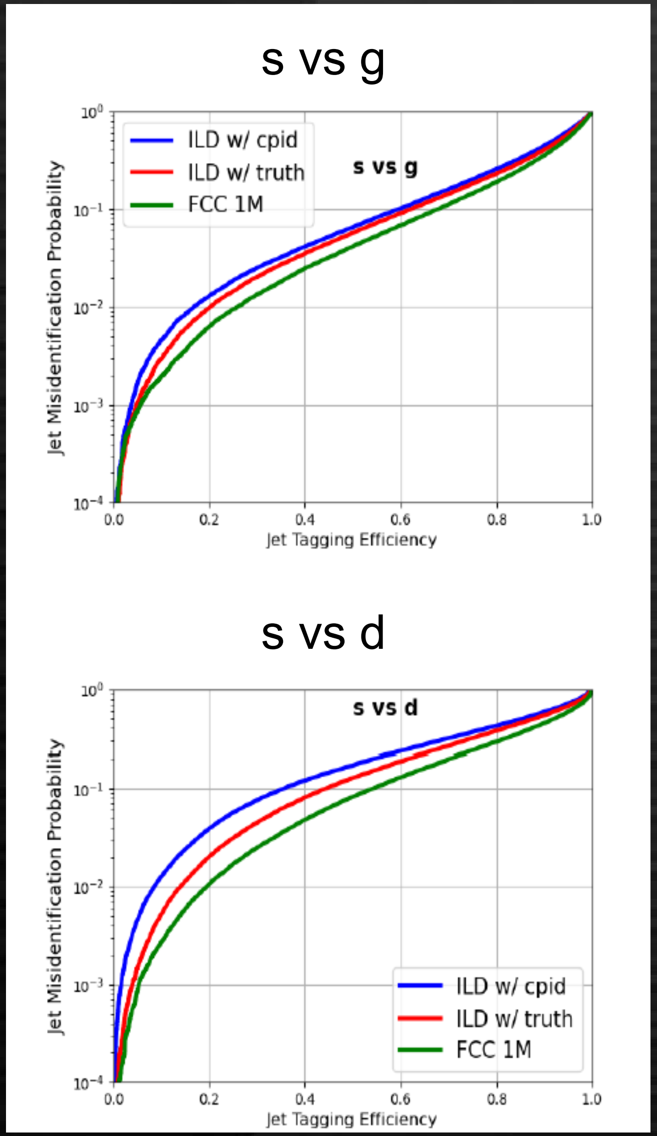
- First results obtained with CPID
  - No significant improvements from old PID: investigating
  - Compared with truth PID: some difference
  - FCC (1M) better than ILD Truth PID
    - Reason needs to be investigated (maybe non-perfect assignment of truth PID)
- Still needs study

	s-tag 80% eff.	
Method	g-bkg acceptance (%)	d-bkg acceptance (%)
ILD full sim. CPID	25.7	42.7
ILD full sim. Truth PID	23.2	38.0
FCC 1M (PID+tof)	20.3	29.6

Strange tagging performance



FCCee plot (in their study)



Applying to H → ss analysis…

- Initial performance obtained: maybe not optimal
  - Difference needs to be understood
    - With truth PID – finite PID performance?
    - With FCCee – truth assignment? Or else
- Implementation to ILCSoft framework ongoing
  - Initial implementation on LCFIPlus
    - Using ONNX output feature in weaver and adapter in FCCee software
    - Gave encouraging results for b/c tagging, checking s tag
    - Quantitative performance confirmation in ~next week

Higgs to ss study: status

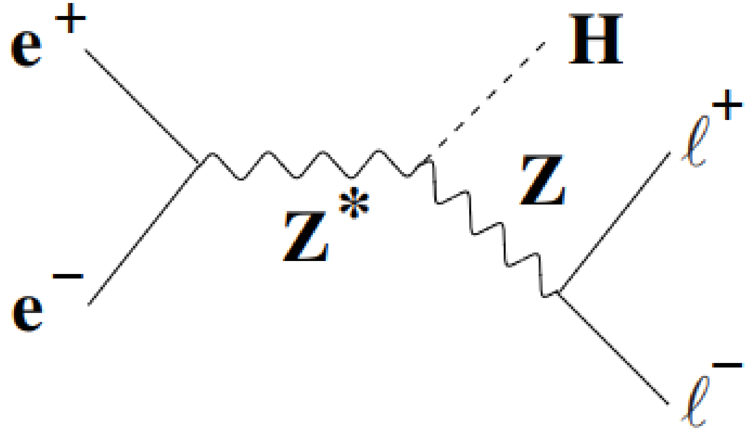
Ryuki Sugawara, Ritsuya Hosokawa, Shinya Narita (Iwate U.)
Taikan Suehara (ICEPP, U. Tokyo)

Objective.

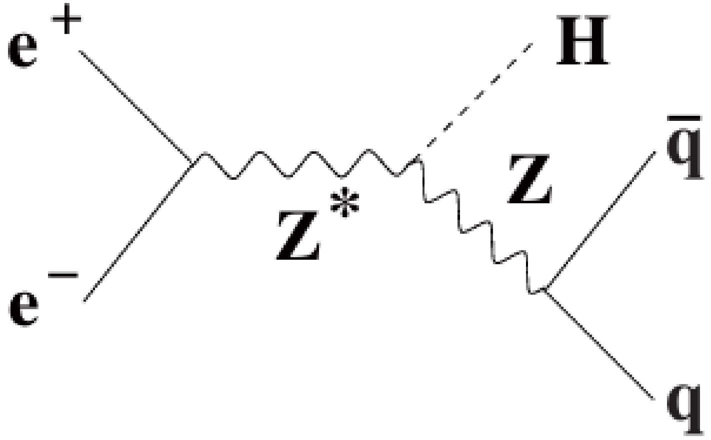
In this study, we estimate the cutting conditions for the Higgs-generated events and background events simulated in the International Linear Collider Experiment.

In addition, the process of background decays to leptons and hadrons is analyzed.

The analysis was performed with reference to H.Ono「Evaluation of measurement accuracies of the Higgs boson branching fractions in the International Linear Collider」



Leptonic Channel signal event



Hadronic Channel signal event

Leptonic channel Analysis Process

After finding the lepton pair, the remaining particles are reconstituted in two jets (2 leptons + 2 jets).

① Find lepton pairs using IsolatedLeptonTagging Processors (signal event)

② Except for the lepton pair I found. Then jetclustering with 2 jets

③ Use MakeClass to apply cuts

④ Perform steps ① ~ ③ again, adding BackGround as well.

Signal Event Result

Results were normalized using cross section and luminosity.

$$Normalized\ Events = Events \times \frac{CrossSection}{Generated} \times Luminosity$$

Change in the number of events when cuts are made to signal events

cut name	reference	my data	normalized
Generated	2917	8219	2717
track ID	2668	7682	2539
Di-lepton mass (GeV)	2287	6538	2161
Z direction	1889	5350	1768
Di-jet mass (GeV)	1445	2401	793
Recoil mass (GeV)	1365	2107	696

cross-section:
10.8691 fb

The signal events did not differ greatly from the reference, but differences were observed in the Di-Jet Mass and Recoil Mass.

- Muon identification by calorimeter information
- Cutting of lepton masses to match Z masses
- Cutting in the Z direction for BGremoval of bosons
- Cutting jets to match Higgs mass
- Cut recoil masses to lepton pairs

These cuts were made for muon events

About the Background

The main background for lepton events is Z boson or W boson-derived events.
The following four events were used in the analysis.

- ZZ_semileptonic
- ZZ_leptonic
- WW_semileptonic
- WW_leptonic

semilptonic : Events contain one lepton pair
leptonic : Events contain two lepton pairs

BackGround Events Result

$$Normalized\ Events = Events \times \frac{CrossSection}{Generated} \times Luminosity$$

This one does not deviate greatly from the reference as well.
However, Di-Jet Mass and Recoil Mass differ slightly from the reference.

Normalized						
cut name	reference	ZZsemi	WWsemi	ZZlepton	WWlepton	total
Generated	45122520	209519	4694775	22239	390855	5317389
track ID	28175	19749	443	1861	4834	26889
Di-lepton mass (GeV)	12901	12277	104	1323	1441	15146
Z direction	8036	7620	78	509	995	9204
Di-jet mass (GeV)	1955	375	0	8	0	383
Recoil mass (GeV)	983	303	0	5	0	308
cross section		838.079	18779.1	88.9574	1563.42	

Hadronic channel Analysis Process

①Use SelectReconstructedParticle(Processor) for signal events to calculate minimum momentum.

②Use ThrustReconstruction(Processor) to calculate thrust.

③Jet clustering with 4 jets.

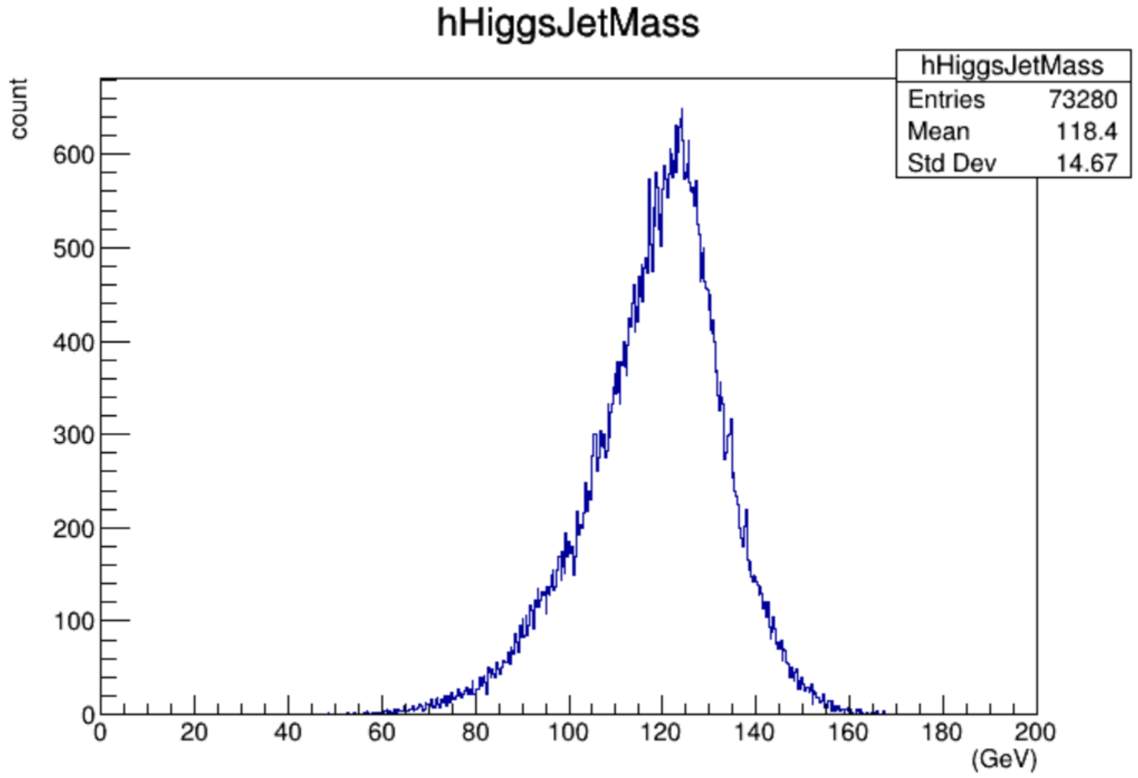
④Analyze in the same way in the background.

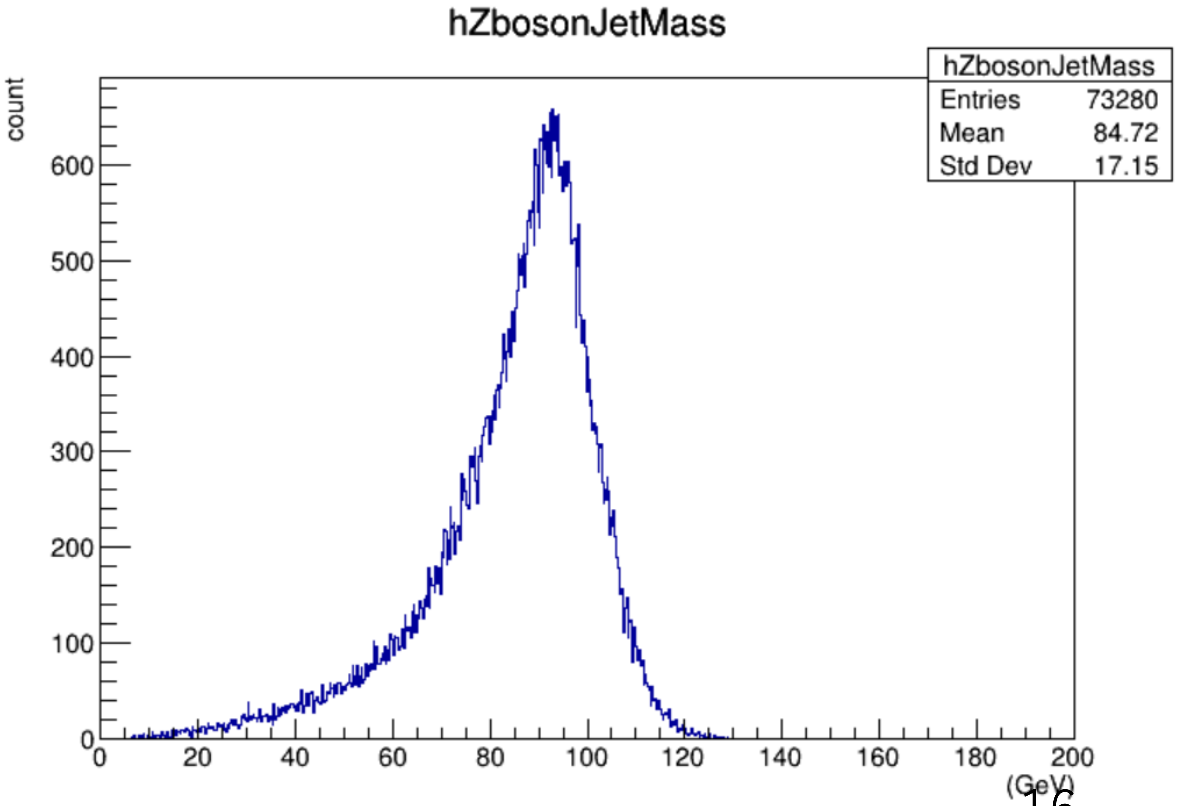
4-jet sorting(signal)

The jets were chosen so that χ^2 in this equation is the smallest.

$$\chi^2 = \left(\frac{M_{j_1 j_2} - M_Z}{\sigma_Z}\right)^2 + \left(\frac{M_{j_3 j_4} - M_H}{\sigma_H}\right)^2,$$

※$M_Z = 91.2\ GeV$, $M_H = 125\ GeV$
$\sigma_H = 4.4\ GeV$ $\sigma_Z = 4.7\ GeV$





Signal Event Result

Results were normalized using cross section and luminosity.

$$Normalized\ Events = Events \times \frac{CrossSection}{Generated} \times Luminosity$$

Change in the number of events when cuts are made to signal events

	reference	my data	normalized
generated	52507	75149	85757
$\chi 2$	32447	33806	38578
charge tracks	25281	33806	38578
Y value	25065	33673	38426
thrust	24688	33159	37839
thrust angle	21892	29137	33250
Z di-jet mass	16359	22109	25230
H di-jet mass	16359	17480	19947
		crossSection : 343.03	

More signal events remained in my data than in the reference.

About the Background

The main background of hadronic events are Z boson or W boson-derived events.
The following two events were used in the analysis

○4f_hadronic
- ZZ_hadronic
- WW_hadronic
- ZZWWMix_hadronic

○2f_hadronic

BackGround Events Result

$$Normalized\ Events = Events \times \frac{CrossSection}{Generated} \times Luminosity$$

More background remained than references.In particular, the cutting of charge tracks does not seem to have been done well.

	reference	4f normalized	2f normalized	total
generated	45904900	3716600	31991500	35708100
$\chi 2$	268980	850391	949192	1799583
charge tracks	1120950	850391	949192	1799583
Y value	1002125	844053	589446	1433499
thrust	935950	839494	453161	1292655
thrust angle	696201	627746	323115	950861
Z di-jet mass	411863	475288	229700	704988
H di-jet mass	411863	355931	169555	525486
cross section		14866.4	127966	

Plans & Timelines

- Finalize strange tagging accessible from ILCSoft
  - Hopefully early next week (now validating the performance)
- Apply strange tagging to a channel to get H→ss performance
  - Maybe leptonic channel
    - Hadronic channel should be more difficult
  - Optimize cuts (high purity is needed for H → ss studies)
  - First result hopefully by 20th Oct.
- Cover most of channels by end of this year
  - llH, qqH, $\nu\nu$H
  - Doing some ML-based selection as well?
- Dependence on detector performance will be studied later