AFB studies at 500 GeV

LCFI+ Flavour Tag Optimization

ILD Top/HF group meeting
3/03/23

Jesús P. Márquez Hernández

February updates

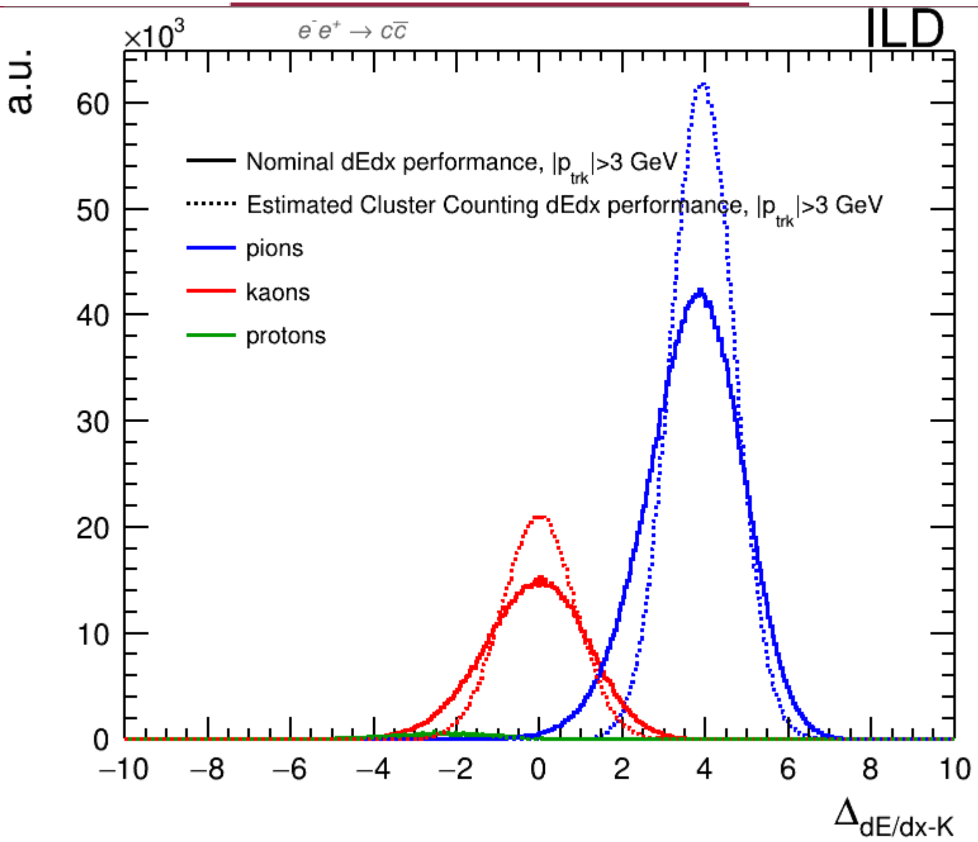
1. Introducing dEdx to LCFI+.
   - Implementation finished and review.
     - Already prepared to be *pulled in Git* :)

2. Re-training of flavour tag weights using dEdx.
   - With a PSO for each case:
     - 250 GeV + dEdx
     - 500 GeV + dEdx
     - 250 GeV + dEdx (+25% improvement)
     - 500 GeV + dEdx (+25% improvement)

Prospects

If we don’t redo the PSO with every new set of data we may run into under-fitting

Prospects for improving dEdx



The gaussians represented by each type of particle would be thinner, allowing better classification when we use them for our observables

First test with dEdx

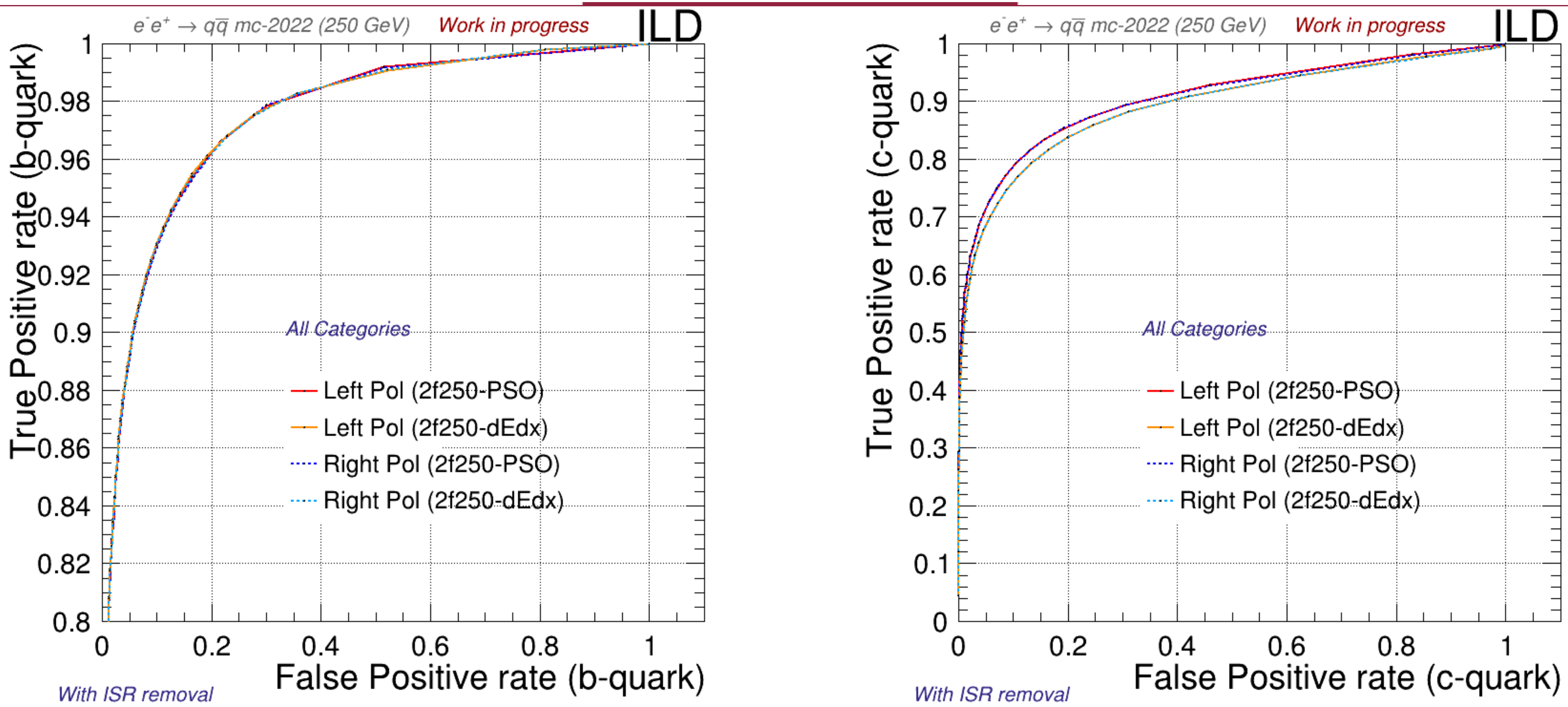
Performance plots

- On the next slides:
  - Plots for b-tag and c-tag:
    - ROC, considering the desired flavour as signal and the others as background.
    - Purity vs Efficiency.
  - Luminosities:

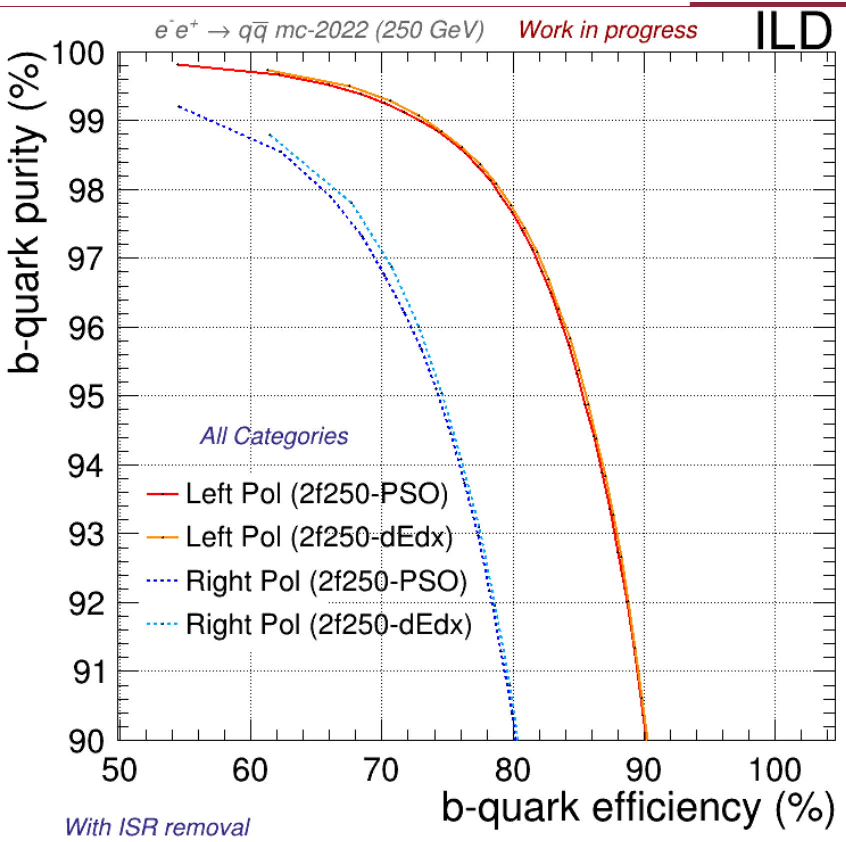
		Test + Training (50/50)	Sample for plots
250 GeV	$e^-_L e^+_R$	20 fb^{-1} (2.5 M events)	26 fb^{-1} (3.3 M events)
	$e^+_R e^+_L$	36 fb^{-1} (2.5M events)	15 fb^{-1} (1.1 M events)

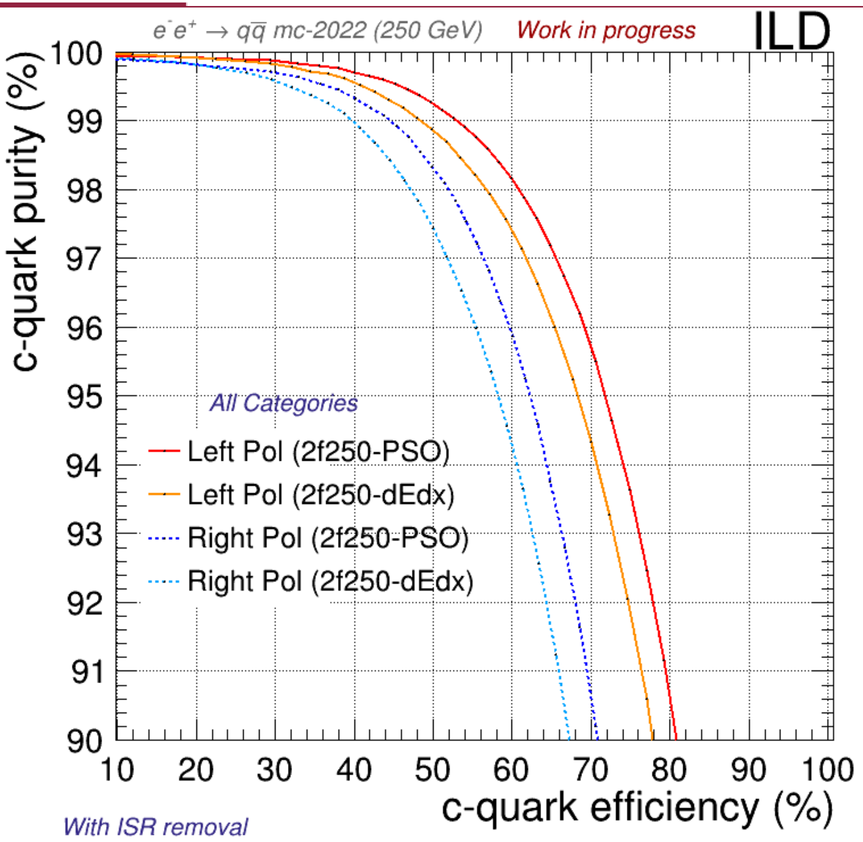
		Test + Training (50/50)	Independent sample for plots
500 GeV	$e^-_L e^+_R$	250 fb^{-1} (1.2M events)	1430 fb^{-1} (7.0 M events)
	$e^+_R e^+_L$	480 fb^{-1} (1.5M events)	1060 fb^{-1} (3.2 M events)

PSO+dEdx Performance (250 GeV)

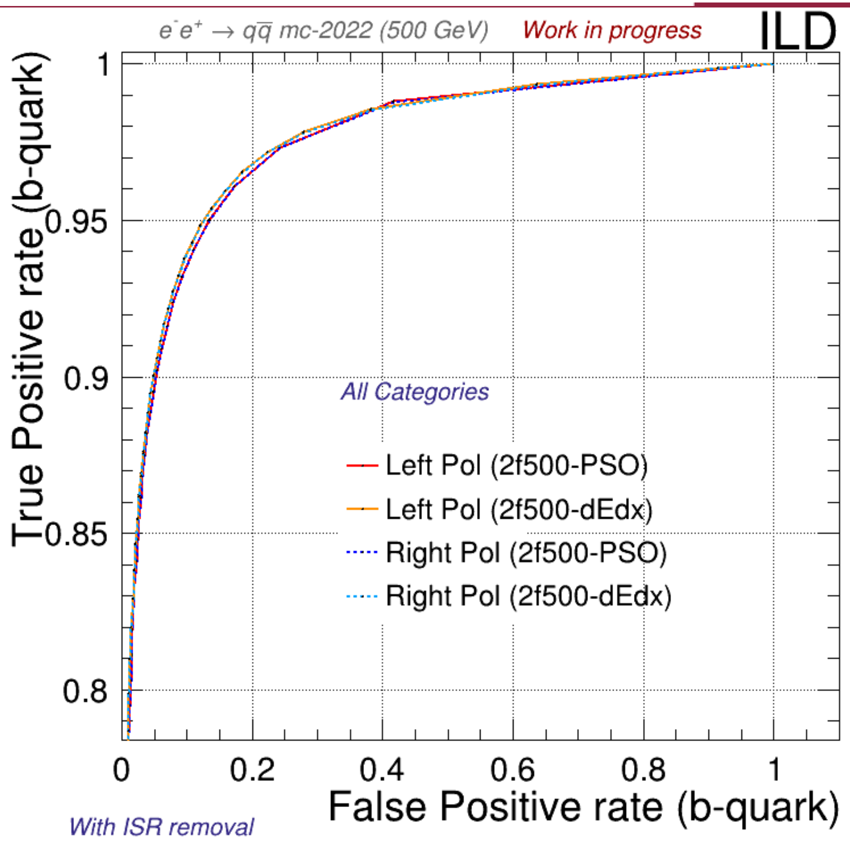


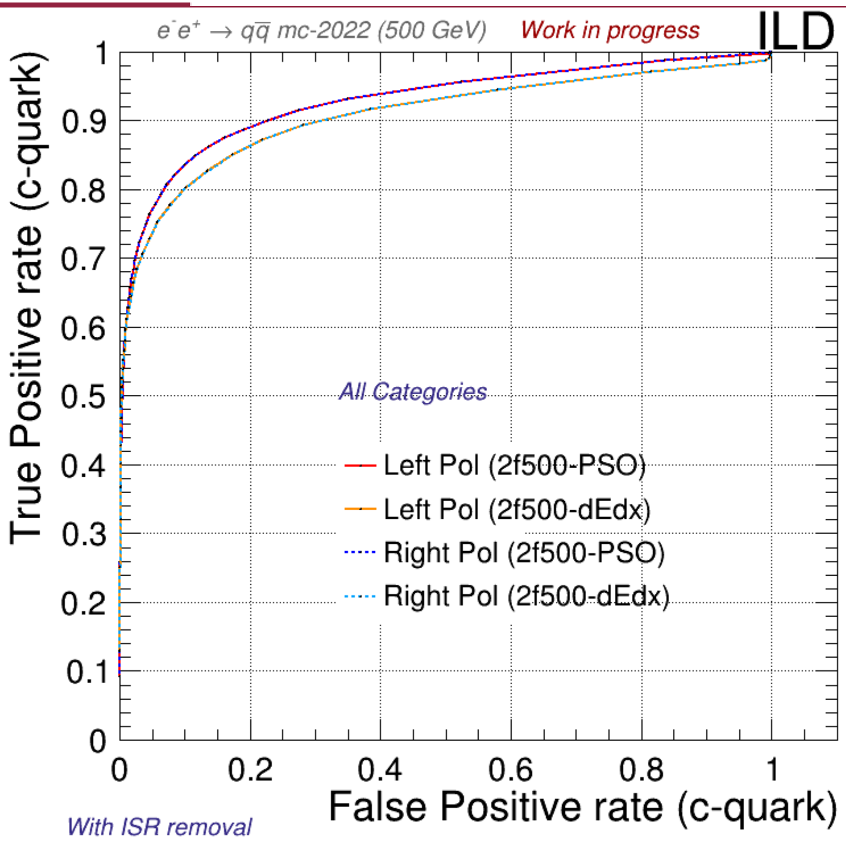
PSO+dEdx Performance (250 GeV)



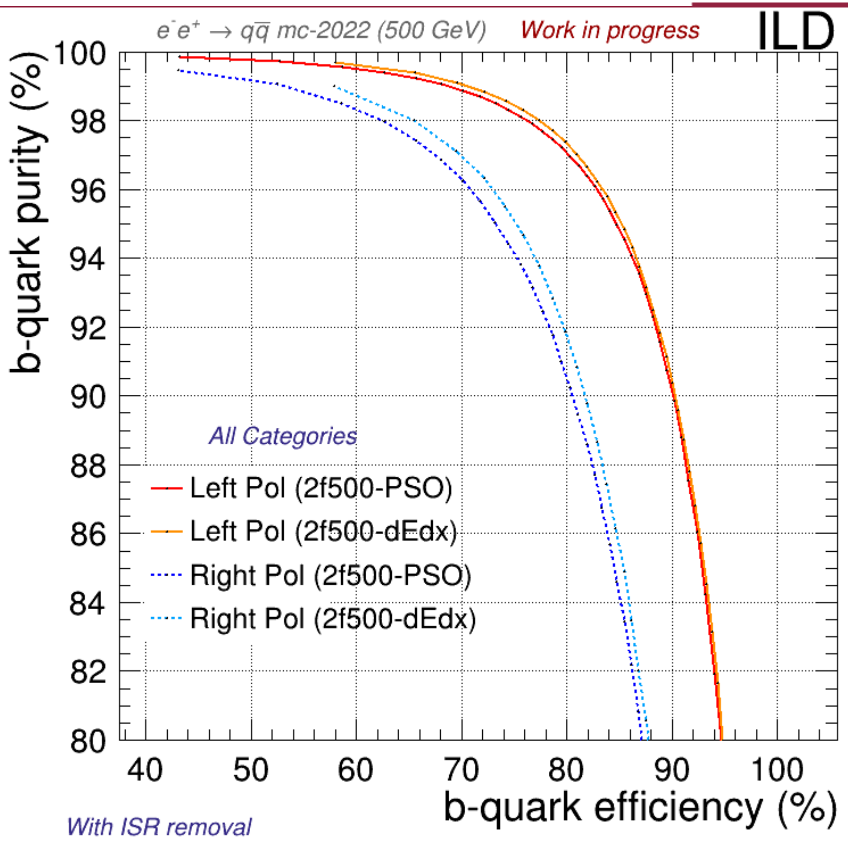


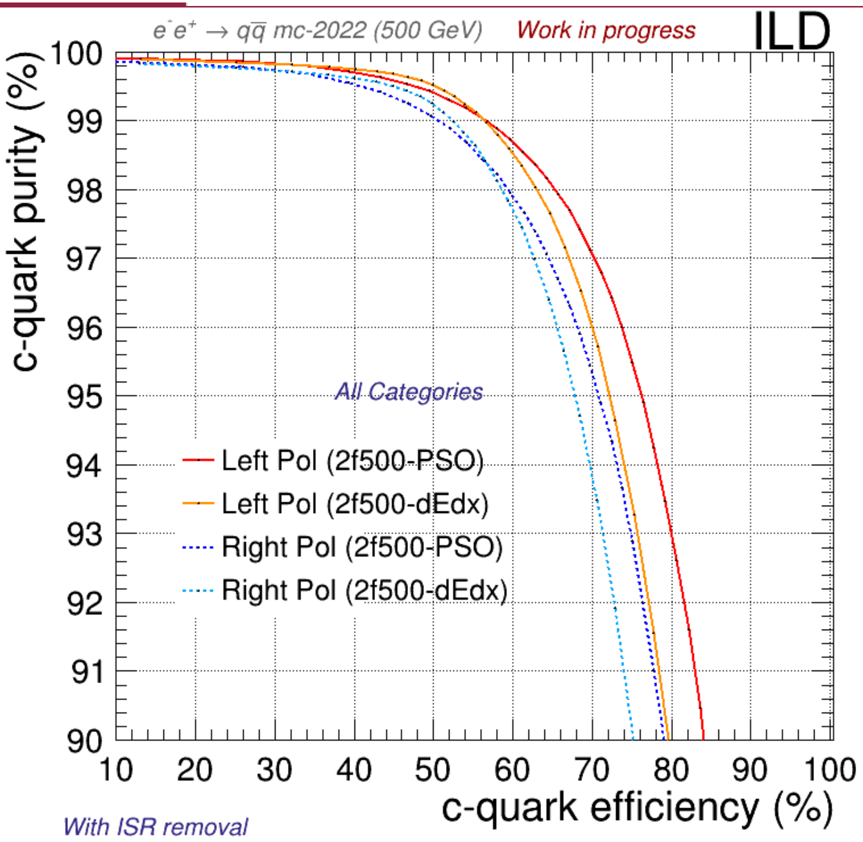
PSO+dEdx Performance (500 GeV)





PSO+dEdx Performance (500 GeV)





dEdx Performance - Insights

- There's some problem in the training process:
  - Maybe some overfitting in catA is causing miss-tagging of c-quarks.
    - After checking it... it seems that most categories are somehow problematic.
      - We might need a stronger statistical tests' thresholds (AD+KS).

- Next move: Repeat the PSO with a higher KS threshold ($p>0.1$).

- Also: Merge the 500 GeV $e_L p_R$ + $e_R p_L$ samples for training.

dEdx Performance - Insights (thresholds)

- For 250 GeV:

	CATEGORY	FOM	AD	KS
PSO	A	0,801	0,300	0,129
	B	0,939	0,169	0,064
	C	0,899	0,271	0,075
	D	0,956	0,974	0,258
PSO + dEdx	A	0,817	0,390	0,058
	B	0,948	0,840	0,170
	C	0,915	0,391	0,160
	D	0,954	0,264	0,080

- For 500 GeV:

	CATEGORY	FOM	AD	KS
PSO	A	0,805	0,518	0,123
	B	0,933	0,101	0,050
	C	0,905	0,276	0,123
	D	0,954	0,509	0,118
PSO + dEdx	A	0,815	0,412	0,056
	B	0,948	0,717	0,130
	C	0,914	0,345	0,141
	D	0,954	0,253	0,077

First test with dEdx

Next training

- For the next training processes:
    - Same structure, just increasing KS threshold to 0.1.
    - Luminosities:

$20\ fb^{-1}$ (2.5 M events) $26\ fb^{-1}$ (3.3 M events)

$e^-_L e^+_R$

250 GeV Test + Training (50/50) Sample for plots

$e^+_R e^+_L$

$36\ fb^{-1}$ (2.5M events) $15\ fb^{-1}$ (1.1 M events)

$250\ fb^{-1}$ (1.2M events) $1430\ fb^{-1}$ (7.0 M events)

$e^-_L e^+_R$

500 GeV Test + Training (50/50) Independent sample for plots

$e^+_R e^+_L$

$480\ fb^{-1}$ (1.5M events) $1060\ fb^{-1}$ (3.2 M events)

Current State (250 GeV)

	CATEGORY	FOM	AD	KS	PSO	RETRAINING	NTuples+Plots
PSO	A	0,8013	0,3000	0,1290	done	done	done
	B	0,9390	0,1692	0,0640	done	done	done
	C	0,8989	0,2711	0,0750	done	done	done
	D	0,9560	0,9743	0,2580	done	done	done
PSO + dEdx	A	0,8240	0,2829	0,1030	done	ongoing	tbd
	B	0,9464	0,3453	0,2110	done	ongoing	tbd
	C	0,9163	0,6668	0,1530	done	ongoing	tbd
	D	0,9604	0,7050	0,4500	done	ongoing	tbd
PSO + dEdx (25% improved)	A				ongoing	tbd	tbd
	B				ongoing	tbd	tbd
	C	0,9190	0,6543	0,1240	ongoing	tbd	tbd
	D	0,9609	0,7213	0,2120	ongoing	tbd	tbd

Current State (500 GeV)

	CATEGORY	FOM	AD	KS	PSO	RETRAINING	NTuples+Plots
PSO	A	0,8053	0,5176	0,1230	done	done	done
	B	0,9328	0,1009	0,0500	done	done	done
	C	0,9049	0,2756	0,1230	done	done	done
	D	0,9540	0,5088	0,1180	done	done	done
PSO + dEdx	A	0,8149	0,4122	0,0600	ongoing	tbd	tbd
	B	0,9479	0,7165	0,1300	ongoing	tbd	tbd
	C	0,9143	0,3451	0,1400	ongoing	tbd	tbd
	D	0,9543	0,2527	0,0800	ongoing	tbd	tbd
PSO + dEdx (25% improved)	A				tbd	tbd	tbd
	B				tbd	tbd	tbd
	C				tbd	tbd	tbd
	D				tbd	tbd	tbd

Summary & Prospects

- First, introducing dEdx to flavour tagging was sucesful, now we have to use it **well**.
- Overfitting problem appeared with the new dEdx variables:
  - KS test's threshold made stronger to avoid this.
- The PSO for different sets of data [old, dEdx, dEdx(+0.25)] is running!

- After completing the production of Flavour Tagging weights we will compare:
  - For **2f250**, PSO+dEdx, PSO+dEdx(+0.25).
  - For **PSO**, PSO+dEdx, PSO+dEdx(+0.25)
- Then, prepare the NTuples for:
  - **physical studies** (R_q & A_{FB})!

Thanks for your attention

Back-up

Events for each category

With ISR removal

Z-Pole (LCFI+ paper[1])

Events (%)			
Cat.	b jets	c jets	uds jets
A	22.9	59.5	98.1
B	39.7	39.8	1.80
C	13.5	0.54	0.02
D	23.8	0.19	0.04

1. LCFIPlus: A Framework for Jet Analysis in Linear Collider Studies

250 GeV samples

Events (%)			
Cat.	b jets	c jets	uds jets
A	13.9	46.2	98.2
B	30.5	51.0	1.59
C	23.9	2.29	0.11
D	31.7	0.55	0.14

500 GeV samples

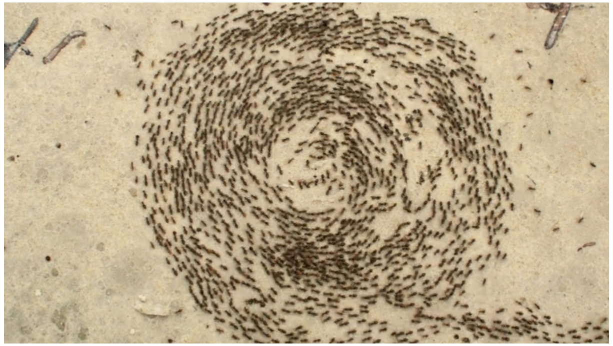
Events (%)			
Cat.	b jets	c jets	uds jets
A	11.2	35.8	96.7
B	28.6	58.3	2.64
C	22.9	4.65	0.26
D	37.3	1.27	0.42

Category	A	B	C	D
Number of vertices	0	1	1	2
Number of single-track pseudovertices	0-2	0	1	0

PSO – Adaptation to FT

- We need:
  - A 3-class classifier (b quarks, c quarks, uds quarks).
  - We also want to avoid overfitting:
    - Kolmogorov-Smirnov test
    - Anderson-Darling test

    Control biased test scores. (more info in back-up)
    Each of them have flaws, so using both is a safer way to go!
  - We need a FOM adapted to 3 different classes.
  - A final check is **always needed**:



Trial and error can go wrong sometimes!

PSO – Kolmogorov-Smirnov Test

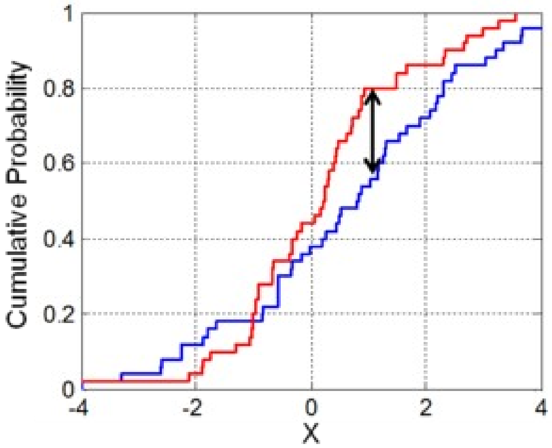
- Compare how likely there is that two different empirical distributions (histograms) came from the same underlying distribution function.
  - It uses the max. distance between the cumulative probability(CPD) of both histograms:

    $$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

  - Then, we past a test for such distance to a certain degree of significance level α (usually 0.05):

    $$D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1+\frac{m}{n}}{2m}}.$$

  - The output is a p-value which determine how likely it is that both histograms came from the same distribution according to our significance level. In particular 1-p is the probability of the null hypothesis (in this case, being different histograms). A 2-sigmas effect is about p < 0.05.



Notice how a big jump in the CPD even in a very narrow region will lead to a very high distance (low KS score): Hyper-sensibility if the distributions are not smooth enough

PSO – Anderson-Darling Test

- The AD test statistic is defined as:

$$A^2 = -n - S$$

- Where:

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln(F(Y_i)) + \ln(1 - F(Y_{n+1-i}))\right]$$

- Being F the cumulative probability distribution for a certain distribution (or the other sample in our 2-samples scenario).
  - Works better with uniform distributions and higher binning.
- Again, the output is an estimator based on a cut in $A > A_{critical}$

Notice how this kind of testing avoid the hyper-sensibility that we had in narrow jumps in the CPD but what if one of these jumps in CPD is actually relevant?

I chose very conservative (and secure) way to proceed: Applying both tests!

PSO – KS & AD Tests in ROOT's histos

- Both test are implemented in ROOT's TH1 Class.
  - We compare test & train TH1 histograms for our study.
  - Very sensitive to binning thickness and doesn't work that good if the underlying distributions are not "smooth" or if the statistics are too small.
  - KS seems to be overly sensitive in older versions of ROOT (v < 6.2 aprox), which we used in ILCSOFT for all this optimization.
- We run it with the "X" function:
  - Run the pseudo experiments post-processor with the following procedure: make pseudo-experiments based on random values from the parent distribution, compare the KS distance of the pseudo-experiment to the parent distribution, and count all the KS values above the value obtained from the original data to Monte Carlo distribution. The number of pseudo-experiments nEXPT is currently fixed at 1000. The function returns the probability estimated as the ratio of pseudo-experiments that pass the test.

We will fix the KS & AD thresholds *ad hoc*, prior inspection, for each category! (example in next slide)