Note

Feng-Yang Hsieh

1 CWoLa

The Classification Without Labels (CWoLa) is a weakly supervised learning method. The CWoLa approach trains a model to discriminate the mixed samples, which are mixtures of the original signal and background samples. The optimal classifier in the CWoLa approach is also the optimal classifier in the traditional fully supervised case where all label information is available. This section utilizes the CWoLa approach to train classifiers on di-Higgs samples.

1.1 Sample

This exercise's signal corresponds to the resonant Higgs boson pairs production in the four-b quarks channel. These Higgs boson pairs are produced via gluon-gluon fusion in the two Higgs doublet model (2HDM). The Higgs boson h ($m_h = 125$ GeV) pair is produced by the heavy CP-even scalar H with mass m_H ranging from 300 GeV to 1200 GeV. The background consists of QCD multi-jet events.

The CWoLa training samples M_1 and M_2 are the mixtures of the signal and background samples. The probability distribution of the mixed sample is a combination of the signal $p_s(x)$ and background $p_B(x)$ distributions:

$$\begin{aligned} p_{M_1}(x) &= f_1 p_S(x) + (1 - f_1) p_B(x) \\ p_{M_2}(x) &= f_2 p_S(x) + (1 - f_2) p_B(x) \end{aligned} \tag{1}$$

where f_1, f_2 are the signal fractions, and x represents the observables used for the classification task.

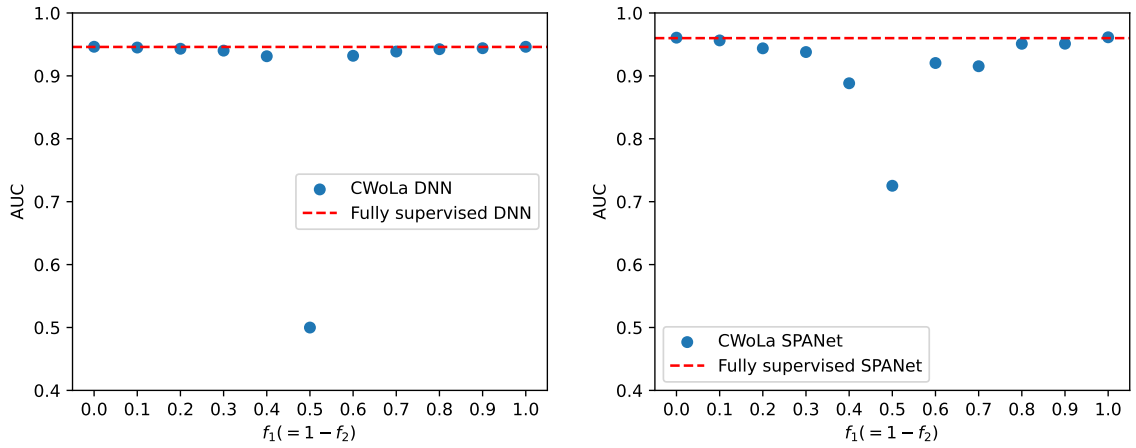
DNN and SPANet network architectures are considered in this exercise. For DNN, the input features are summarised in Table 1, consisting of 16 variables. For SPANet, the input features are a list of final jets, each represented by their 4-momentum $(p_{\mathrm{T}}, \eta, \phi, M)$ and a boolean b-tag.

Table 1: Input variables used to train the dense neural network.

Reconstructed objects	Variables used for training	#
Higgs candidate	$(p_{\mathrm{T}}, \eta, \phi, m)$	8
Subjets	$\Delta R(j_1, j_2)$	2
b-tagging	Boolean for $j_i \in h_{1,2}^{\mathrm{cand}}$	4
Di-Higgs system	$p_{\mathrm{T}}^{hh}, m_{hh}$	2

1.2 Result

The CWoLa training utilizes samples with different signal fractions f_1, f_2 to train the classifiers. The results of CWoLa training are shown in Figure 1 with different signal fractions. When f_1 is far from 0.5, the results tend to approach those of the fully supervised case.



(a) DNN (b) SPANet

Figure 1: The AUC of CWoLa training as a function of the signal fraction f_1. For simplicity, we set signal fraction f_2 equal to $1 - f_1$. The horizontal dashed line indicates the fully-supervised AUC.

When $f_1 = 0.5$ the mixed sample M_1 and M_2 have identical distributions, so the classifier can not learn anything in this case. In the case of DNN, the AUC is 0.5, as expected. However, for SPANet, the AUC is more than 0.7.

This is because SPANet is trained on both pairing and classification tasks simultaneously. The pairing part introduces asymmetries between signal and background samples, leading to the AUC that deviates from 0.5.

To investigate the effect of the pairing task on SPANet's performance, the weight of the

pairing component is set to zero, meaning that SPANet focuses solely on the classification task. Figure 2 shows the SPANet training results without pairing task. As expected, the AUC is close to 0.5 when $f_1 = 0.5$.

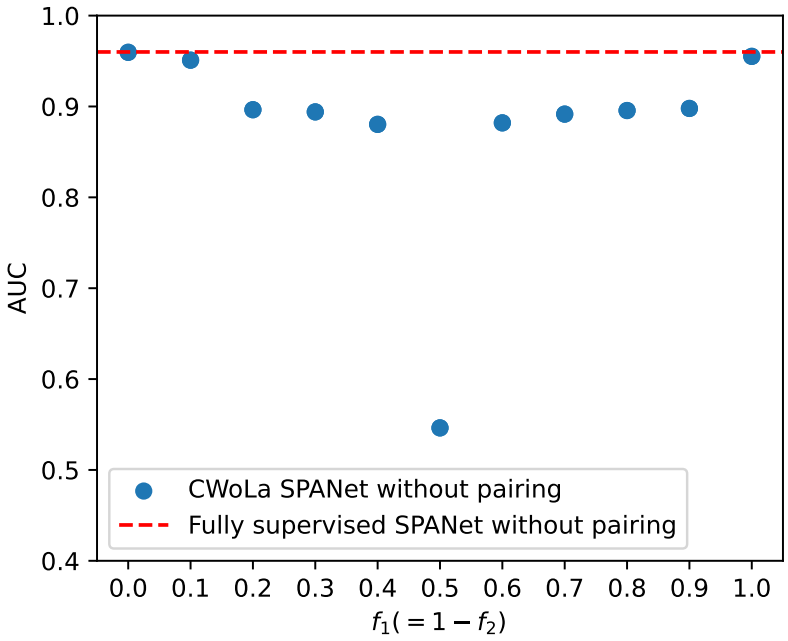


Figure 2: The AUC of CWoLa SPANet training as a function of the signal fraction f_1. For simplicity, we set signal fraction f_2 equal to $1 - f_1$. Here, SPANet is trained only on the classification task.

2 CWoLa hunting

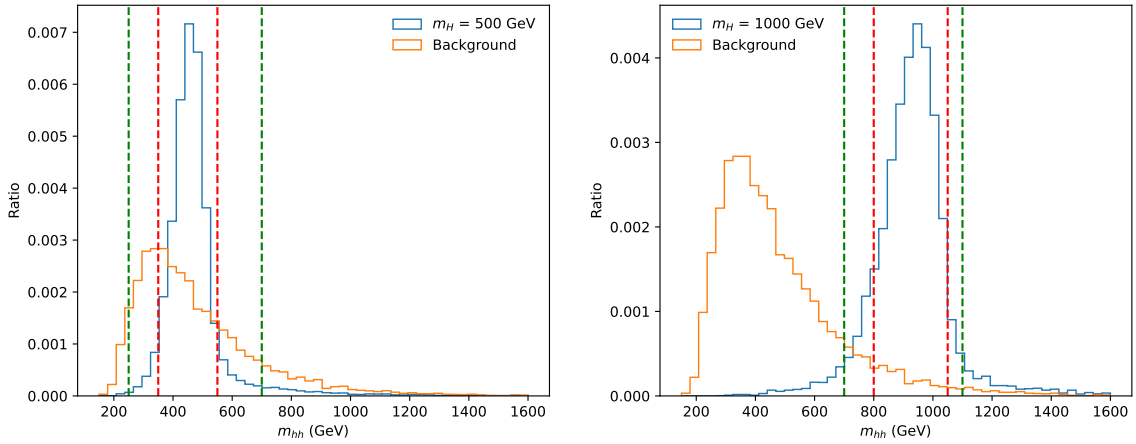
The CWoLa hunting approach considers a m_{res} variable. For background, the m_{res} distribution is smooth while signal m_{res} distribution is expected to be localized near some m_0. Consequently, this variable could be used to create two mixed samples. Additional features that are uncorrelated with m_{res} can be used for training a classifier. This technic is first introduced by Reference [1].

2.1 Sample

The signal is the resonant Higgs boson pairs production in the four-b quarks channel. This section produces the Higgs boson pair by the heavy CP-even scalar H with mass $m_H = 500$ GeV or $m_H = 1000$ GeV. The background consists of QCD multi-jet events. The

basic requirement is the "four-tag cut," which requires at least four b-tagged $R = 0.4$ anti-k_t jets with $p_{\mathrm{T}} > 40$ GeV and $|\eta| < 2.5$. Only the events passing the four-tag cut are used in the following analysis.

The CWoLa hunting approach utilizes the signal and sideband regions to create the mixed training sample. The di-Higgs system's total invariant mass m_{hh} determines the signal and sideband region. This quantity is computed from the four b-jets with the highest transverse momentum. Figure 3 presents the m_{hh} distribution of signal and background samples. Table 2 summarizes the signal and sideband regions. These signal and sideband regions are chosen so that the corresponding cross-sections are closed.



(a) $m_H = 500$ GeV (b) $m_H = 1000$ GeV

Figure 3: The total invariant mass m_{hh} distribution of signal and background samples. The signal region is between the red dashed lines. The sideband region is between the green dashed lines and excludes the signal region.

Table 2: The signal and sideband regions with different resonant samples. The unit is GeV.

m_H	Signal	Sideband
500	$[350, 550]$	$[250, 350] \cup [550, 700]$
1000	$[800, 1050]$	$[700, 800] \cup [1050, 1100]$

Table 3 is the cutflow table of the selection cuts. The number of events used in mixed training samples could be computed from these cross-sections. The training sample size is presented in Table 4.

Consider the DNN CWoLa classifier. The Higgs candidates are reconstructed by the min-ΔR pairing method. In the min-ΔR method, the four b-tagged jets with the highest p_{T}

Table 3: The cross sections for the di-Higgs signal and background processes at different selection cuts.

m_H (GeV)		Cross section (fb) Signal	Background	S/B	$\mathcal{L} = 139$ fb^{-1} $S/\sqrt{B}$
500	Four tag	3.64	6.03e+03	6.03e-04	0.553
	Signal region	3.13	2.57e+03	1.22e-03	0.727
	Sideband region	0.35	2.36e+03	1.50e-04	0.086
1000	Four tag	0.081	6.03e+03	1.34e-05	0.0123
	Signal region	0.063	3.32e+02	1.90e-04	0.0408
	Sideband region	0.010	3.19e+02	3.03e-05	0.0064

Table 4: The training sample size for the mixed sample. The luminosity is $\mathcal{L} = 78$ fb^{-1} because the generated samples are not enough for now.

m_H (GeV)	Mixed sample	True label Signal	Background
500	M_1	244	200k
	M_2	28	184k
1000	M_1	5	26k
	M_2	1	25k

are used to form the two Higgs boson candidates. The min-ΔR method selects the pairing configuration in which the higher-p_{T} jet pair has the smallest ΔR separation. The input features are similar to the previous case (Table 1), but the b-tagging information and the di-Higgs system's total invariant mass are excluded. min-ΔR pairing only uses the b-tagged jets. Total invariant mass is already used to determine the signal and sideband region.

2.2 Training results

Table 5 presents the DNN classification training results. These numbers are evaluated from the pure samples, which consist of 5k signal events and 5k background events. The training datasets with and without signal events have similar results. This suggests that the DNN fails to distinguish the signal and background samples but learns the difference between the signal and sideband region. Moreover, the results also imply the input features may correlate to the total invariant mass of the di-Higgs system.

Table 5: The CWoLa DNN training results. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)		ACC	AUC
500	With signal	0.708 ± 0.002	0.770 ± 0.007
	No signal	0.705 ± 0.003	0.769 ± 0.009
1000	With signal	0.868 ± 0.024	0.925 ± 0.023
	No signal	0.850 ± 0.033	0.909 ± 0.026

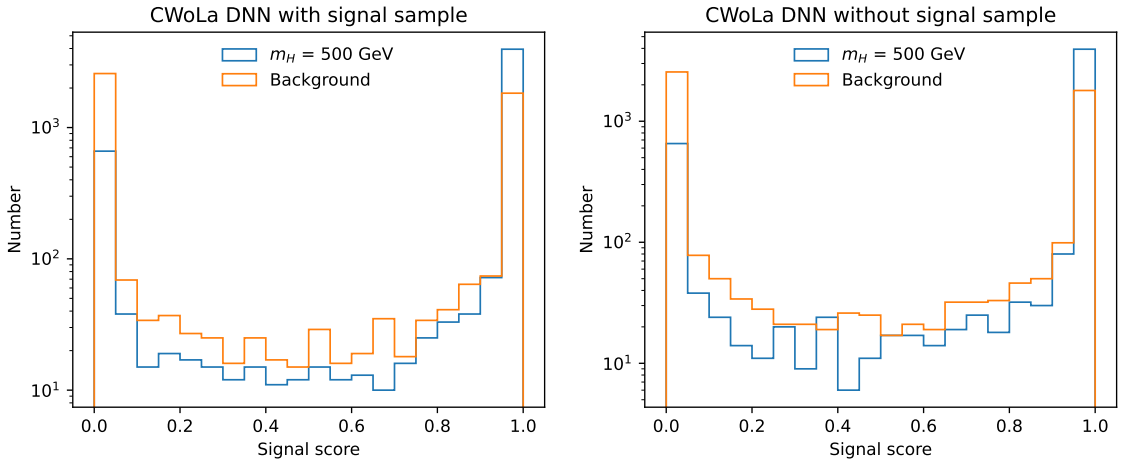
Figure 4 shows the signal score distributions. Even though the signal scores are very different for signal and background distributions, the difference probably stems from the m_{hh} distribution.

There are two issues:

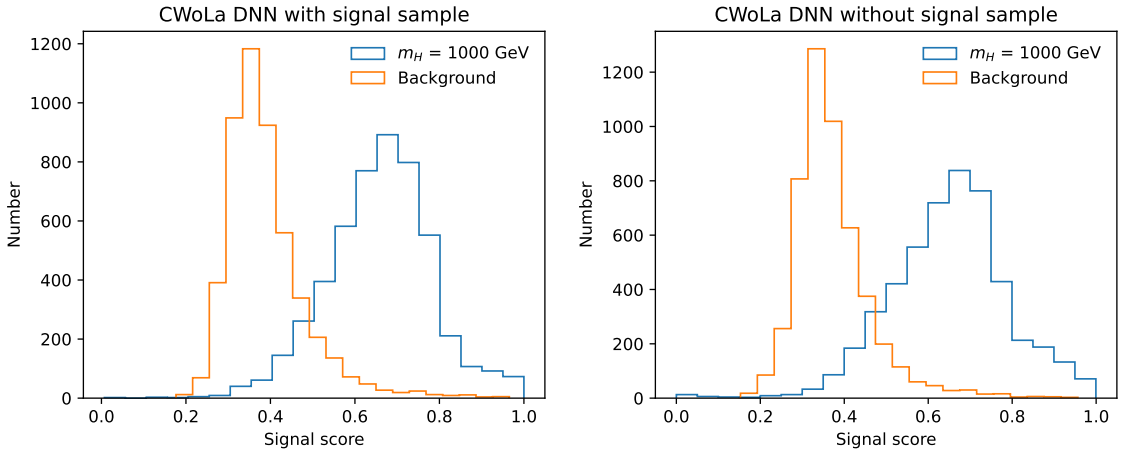
- The input features might correlated to the observables used to determine the signal and sideband region. We need to construct other independent input variables.
- The signal fraction is too low. It is hard to learn something about signal events.

2.3 Correlation matrix

The results in Section 2.2 imply that the di-Higgs system's total invariant mass is not independent of other input features. To find the variables that are highly dependent on



(a) $m_H = 500$ GeV



(b) $m_H = 1000$ GeV

Figure 4: The signal score distributions. We apply the CWoLa DNN on pure samples to obtain the signal score distributions.

the total invariant mass, the correlation coefficients are computed among these variables. Figure 5 and 6 are correlation coefficients on the 500 GeV and 1000 GeV cases, respectively.

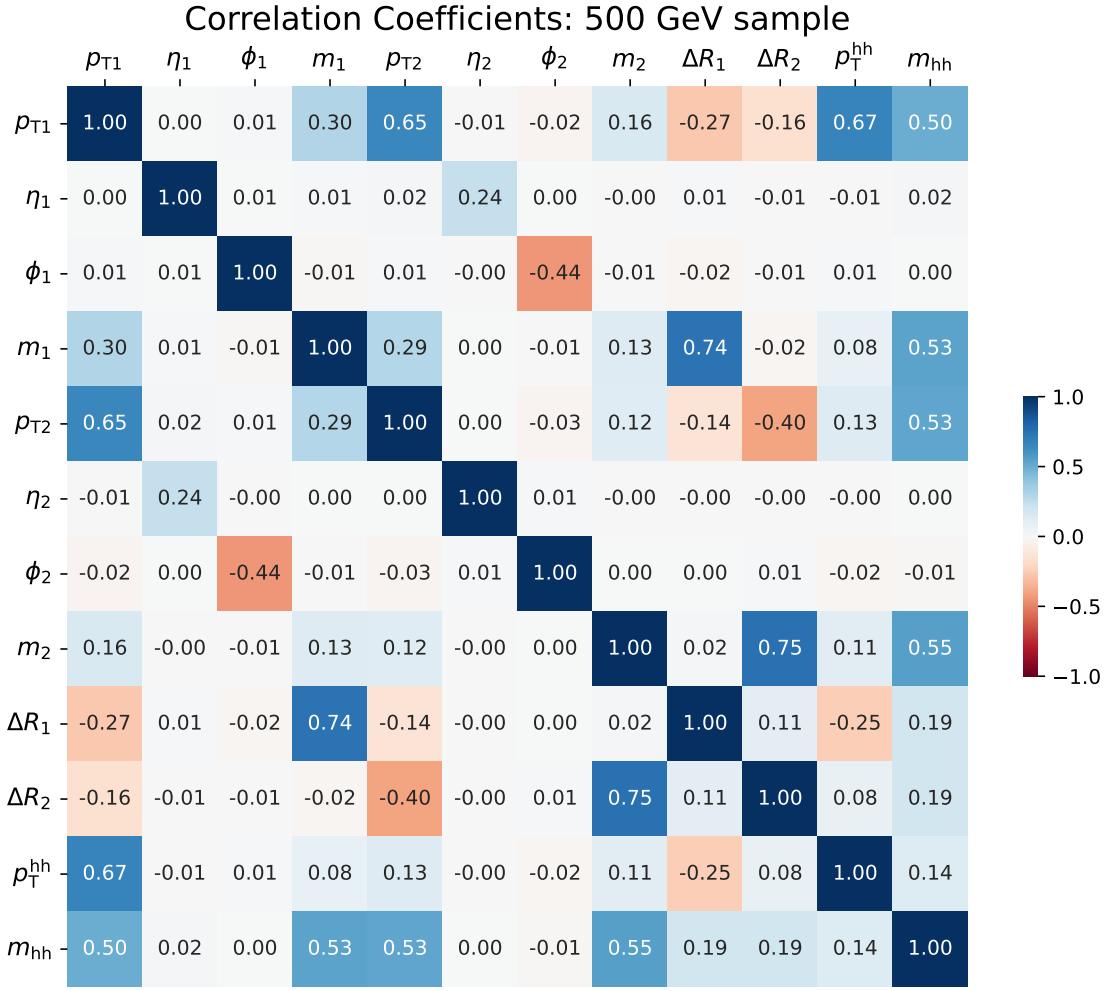


Figure 5: The correlation coefficients among different variables are computed from 500 GeV testing sample, consisting of 5k signal and 5k background.

The results show that the transverse momentum p_T and the invariant mass m of Higgs candidates are highly correlated to the total invariant mass. Figure 7 shows the scatter plots of the transverse momentum of the leading Higgs candidate and the total invariant mass m_{hh}. These plots also explain why the DNN only trained on background samples can distinguish the signal and background events, because the background distribution in the signal and sideband regions are different.

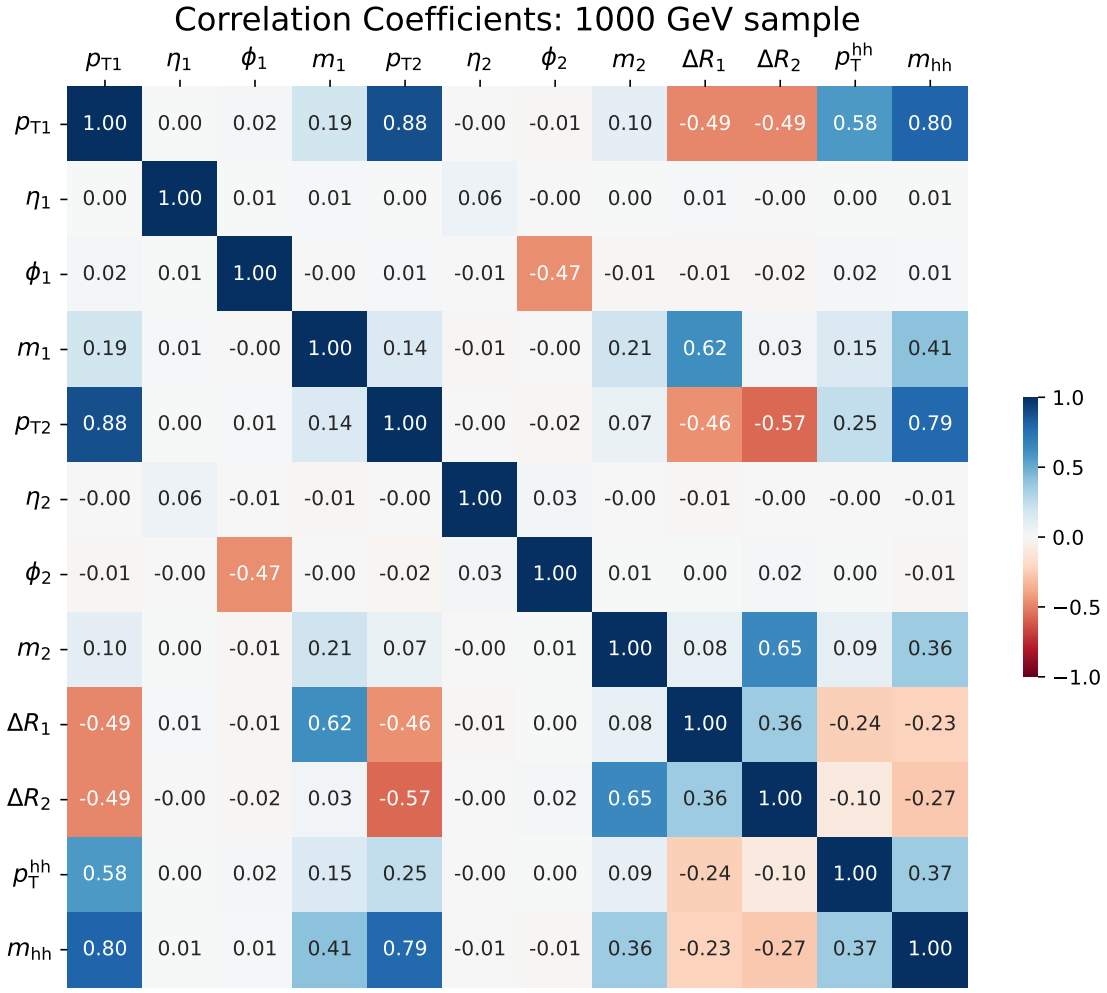
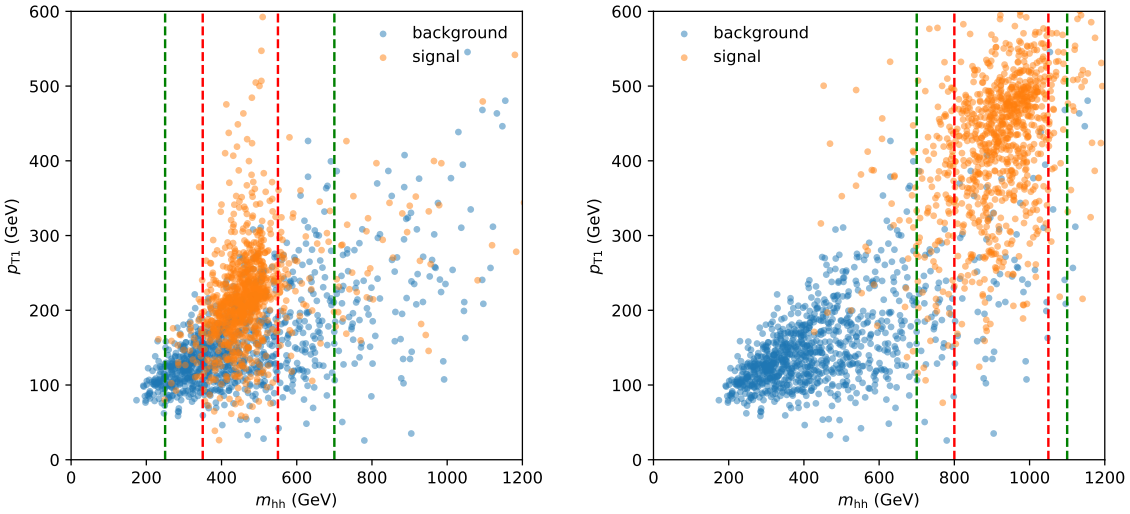


Figure 6: The correlation coefficients among different variables are computed from 1000 GeV testing sample, consisting of 5k signal and 5k background.



(a) $m_H = 500$ GeV (b) $m_H = 1000$ GeV

Figure 7: The scatter plots of the transverse momentum of leading Higgs candidate p_{T1} and total invariant mass m_{hh} distribution. The signal region is between the red dashed lines. The sideband region is between the green dashed lines and excludes the signal region.

2.4 Remove highly correlated features

Figure 5 and 6 show that the transverse momentum p_T and the invariant mass m of Higgs candidates are highly related to the total invariant mass m_{hh}. To investigate the impact of these highly correlated features on the discrimination power of CWoLa DNN models, we remove these input features and train the DNN model again.

Table 6: The CWoLa DNN training results. The transverse momentum and invariant mass of Higgs candidates are removed from samples. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)		ACC	AUC
500	With signal	0.526 ± 0.020	0.536 ± 0.053
	No signal	0.532 ± 0.015	0.543 ± 0.029
1000	With signal	0.586 ± 0.030	0.625 ± 0.046
	No signal	0.564 ± 0.024	0.583 ± 0.042

Table 6 summarizes the results of the CWoLa DNN training without p_T and m features. The training datasets with and without signal events still have similar results. Compared to the previous one (Table 5) the accuracy values are closer to 0.5. These results suggest that

the removed features significantly contribute to the model's discrimination power, and the remaining parameters are hard to utilize to distinguish the signal and background events.

2.5 Transverse momentum cut testing

In Figure 3, the distribution of the background sample exhibits a gradual termination around 150 GeV. To investigate whether this termination is a result of the "four-tag cut", which requires $p_{\mathrm{T}} > 40$ GeV, total invariant mass distributions with different p_{T} cuts are plotted in Figure 8. As the transverse momentum requirement increases from 40 GeV to 70 GeV, the termination point also shifts to larger values. Moreover, the termination remains gradual rather than an abrupt cut-off, suggesting that the gradual termination indeed results from the transverse momentum cut.

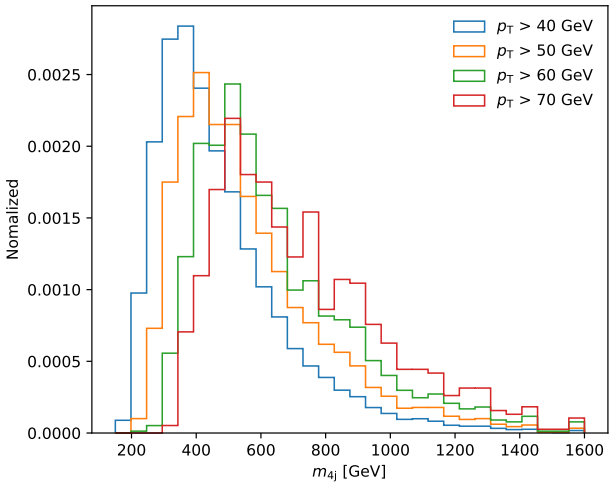


Figure 8: The total invariant mass $m_{4\mathrm{j}}$ distribution of background samples. The transverse momentum requirement varies from 40 GeV to 70 GeV.

2.6 Enlarge the signal sample size

Another issue arises from the low signal fraction (Table 4), making DNN difficult to extract meaningful information about signal events. To investigate the impact of signal sample size, we increase the signal size manually and retrain the DNN model. The training sample sizes are summarized in Table 7.

Table 8 provides the results of the CWoLa DNN training without p_{T} and m features. For the 500 GeV case, the "0 times,", "1 times," and "10 times" samples yield similar results, while "100 times" sample exhibits better performance. This suggests that the CWoLa DNN

Table 7: The training sample size for the mixed sample. Various signal sizes are considered, and the background sizes are fixed for all cases. "1 times" represents the previous "With signal" case and "0 times" represents the previous "No signal" case.

		Signal				Background
m_H (GeV)	Mixed sample	1 times	0 times	10 times	100 times	All
500	M_1	244	0	2438	24380	200k
	M_2	28	0	276	2760	184k
1000	M_1	5	0	49	492	26k
	M_2	1	0	8	75	25k

Table 8: The CWoLa DNN training results. The transverse momentum and invariant mass of Higgs candidates are removed from samples. 1 time and 0 times are the with signal and no signal case in Table 6. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)	times	ACC	AUC
500	1	0.526 ± 0.020	0.536 ± 0.053
	10	0.531 ± 0.027	0.533 ± 0.045
	100	0.634 ± 0.014	0.751 ± 0.030
	0	0.532 ± 0.015	0.543 ± 0.029
1000	1	0.586 ± 0.030	0.625 ± 0.046
	10	0.626 ± 0.027	0.678 ± 0.040
	100	0.621 ± 0.012	0.670 ± 0.023
	0	0.564 ± 0.024	0.583 ± 0.042

can extract meaningful information from the "100 times" sample. In the case of 1000 GeV, we can obtain better results when the signal sample size increases. The performance of 10 times and 100 times is similar. It seems that the training performance is saturated.

Additional samples within this size range are generated to understand further the behavior between 10 times and 100 times samples for the 500 GeV case, and the DNN is trained on these samples.

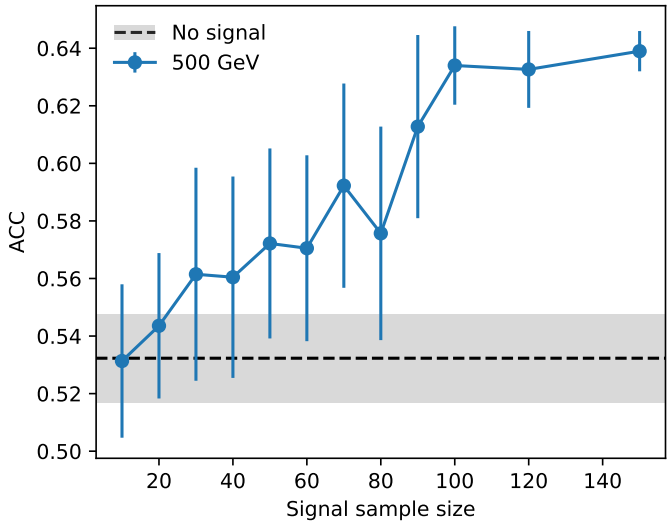


Figure 9: The accuracy of CWoLa DNN training as a function of the signal size. The unit of sample size is the size of the "1 times" case. The error bar is the standard deviation of 10 training. The grey band is the error bar of the "without signal" case.

Figure 9 is the training performance against the signal sample size. In this region, the performance increases when the signal size is increased. 120 times and 150 times samples are also generated and used in training. The accuracy is saturated at around 63%.

Similarly, for the 1000 GeV case, the DNN is trained on samples with sizes ranging from 1 to 10 times. Figure 10 is the training performance against the signal sample size. The performance is similar for all cases. The training accuracy is saturated at around 62%.

2.7 Training with deeper model

In Figure 10, the performance of CWoLa DNN is quickly saturated. To investigate the impact of the model structure, the deeper DNN model is trained. In Section 2.6, DNN consists of 2 hidden layers, while in this section we train the DNN with 4 hidden layers.

The DNN is trained on signal sample size ranging from 1 to 500 times. Table 9 and Figure 11 are the training results. The performance is generally better than the previous

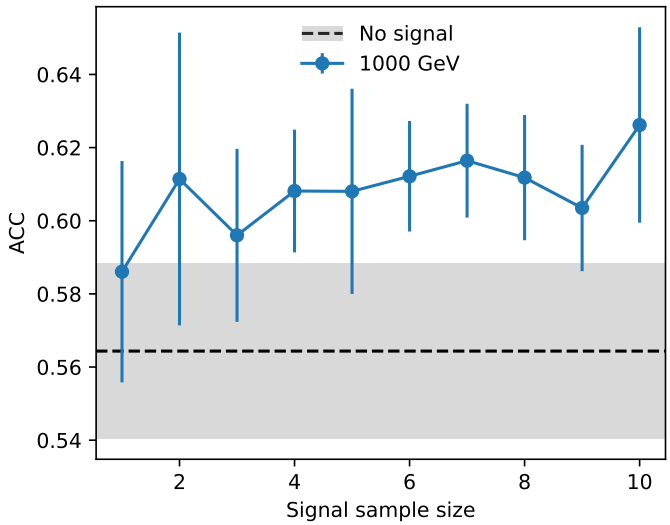


Figure 10: The performance of CWoLa DNN training as a function of the signal size. The unit of sample size is the size of the "1 times" case. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

results (Table 8), even for the without signal case. It seems that the previous model structure is too simple and it limits the training performance. For the signal size within 1 time to 300 times, the performance increases when the signal size increases. After this region, the accuracy does not significantly improve. The accuracy is saturated at around 67%, but this value is still better than the previous ones.

Table 9: The CWoLa DNN training results. The transverse momentum and invariant mass of Higgs candidates are removed from samples. 1 time and 0 times are the with signal and no signal case in Table 6. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)	times	ACC	AUC
1000	1	0.613 ± 0.017	0.649 ± 0.021
	10	0.622 ± 0.018	0.673 ± 0.033
	100	0.639 ± 0.022	0.695 ± 0.034
	0	0.602 ± 0.022	0.643 ± 0.042

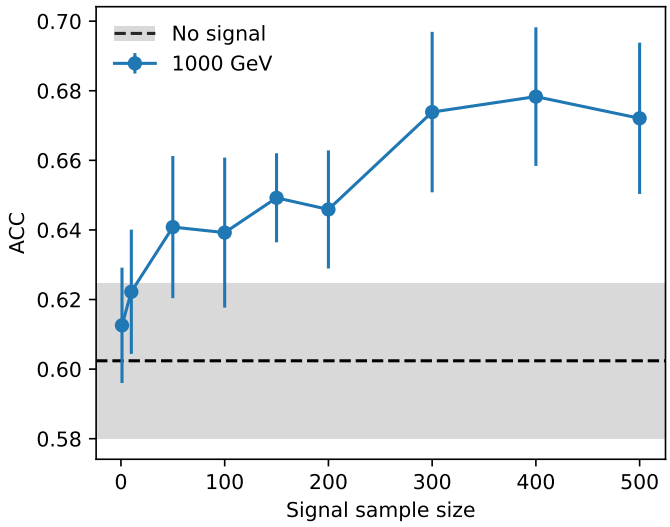


Figure 11: The performance of CWoLa DNN training as a function of the signal size. The unit of sample size is the size of the "1 times" case. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

3 Physical data augmentation

The physical augmentations are inspired by Reference [2], which considers the rotation and smearing augmentations. These augmentations reflect both the symmetries in the physical event and the experimental resolution of the detector.

3.1 Original training data

The signal is the resonant Higgs boson pairs production in the four-b quarks channel. This section produces the Higgs boson pair by the heavy CP-even scalar H with mass $m_H = 500$ GeV. The background consists of QCD multi-jet events. The basic requirement is the "four-tag cut," which requires at least four b-tagged $R = 0.4$ anti-k_t jets with $p_{\mathrm{T}} > 40$ GeV and $|\eta| < 2.5$. Only the events passing the four-tag cut are used in the following analysis.

The training samples consist of 50k signal events and 50k background events and the testing samples consist of 5k signal events and 5k background events.

The Higgs candidates are reconstructed by the min-ΔR pairing method. The input features are similar to the previous case (Table 1), but the b-tagging information is excluded.

3.2 Physical augmentation

We consider three different physical augmentations.

1. Azimuthal rotation: The final state is rotated by an angle ϕ randomly sampled from $[0, 2\pi]$.

2. $\eta - \phi$ smearing: The (η, ϕ) coordinate of Higgs candidates are resampled according to a Normal distribution centered on the original coordinate and with a standard deviation inversely proportional to the p_{T}

$$\eta' \sim \mathcal{N}\left(\eta, \frac{\Lambda}{p_{\mathrm{T}}}\right), \quad \phi' \sim \mathcal{N}\left(\phi, \frac{\Lambda}{p_{\mathrm{T}}}\right) \tag{2}$$

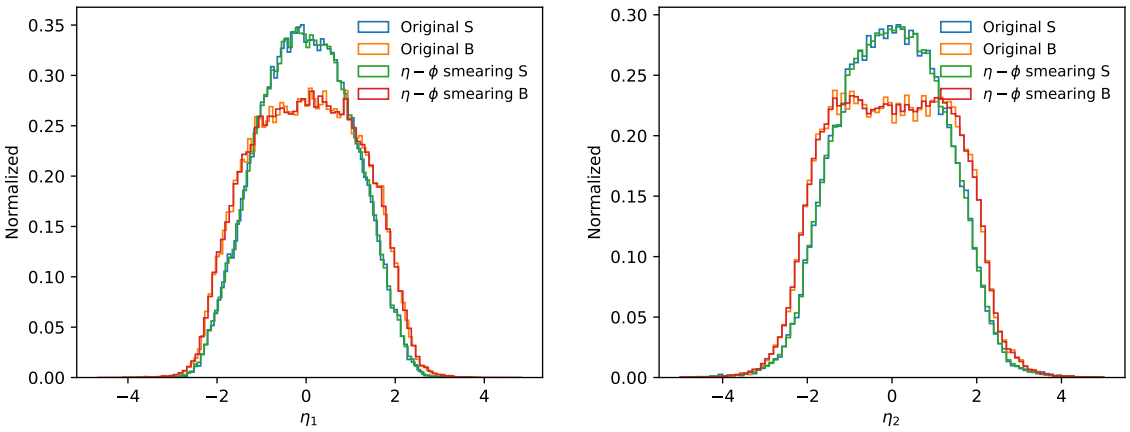
   where η', ϕ' are the augmented coordinate, p_{T} is the transverse momentum of the Higgs candidate, and the smearing scale is set to be $\Lambda = 10$ GeV.

3. p_{T} smearing: The p_{T} of Higgs candidates are resampled according to

$$p'_{\mathrm{T}} \sim \mathcal{N}\left(p_{\mathrm{T}}, f(p_{\mathrm{T}})\right), \quad f(p_{\mathrm{T}}) = \sqrt{0.052 p_{\mathrm{T}}^2 + 1.502 p_{\mathrm{T}}} \tag{3}$$

   where p'_{T} is the augmented transverse momentum, $f\left(p_{\mathrm{T}}\right)$ is the energy smearing applied by `Delphes` (the p_{T}'s are normalised by 1 GeV).

Figure 12, 13 and 14 are the distributions before and after the augmentation. The distributions for the $\eta - \phi$ smearing are similar for both cases. For p_{T} smearing, the peak broadens and the transverse momentum distribution looks smoother.

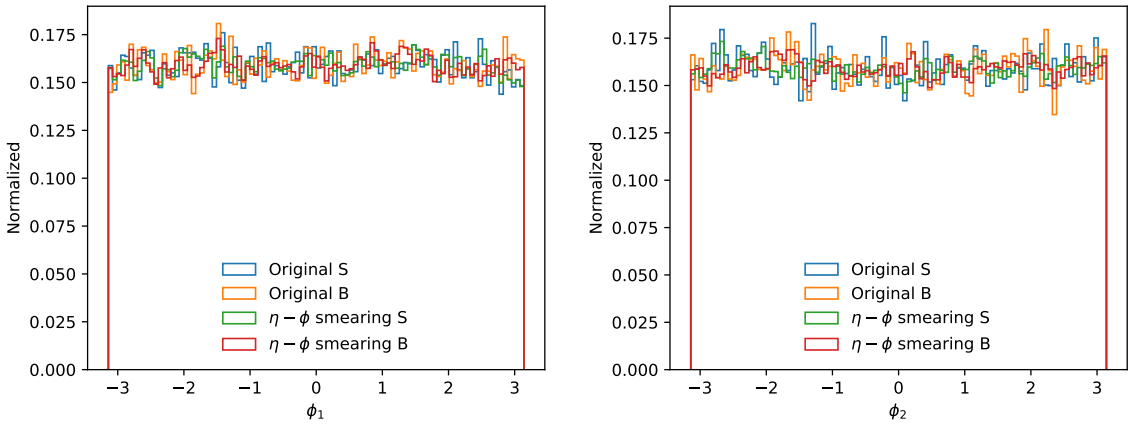


(a) Leading Higgs

(b) Sub-leading Higgs

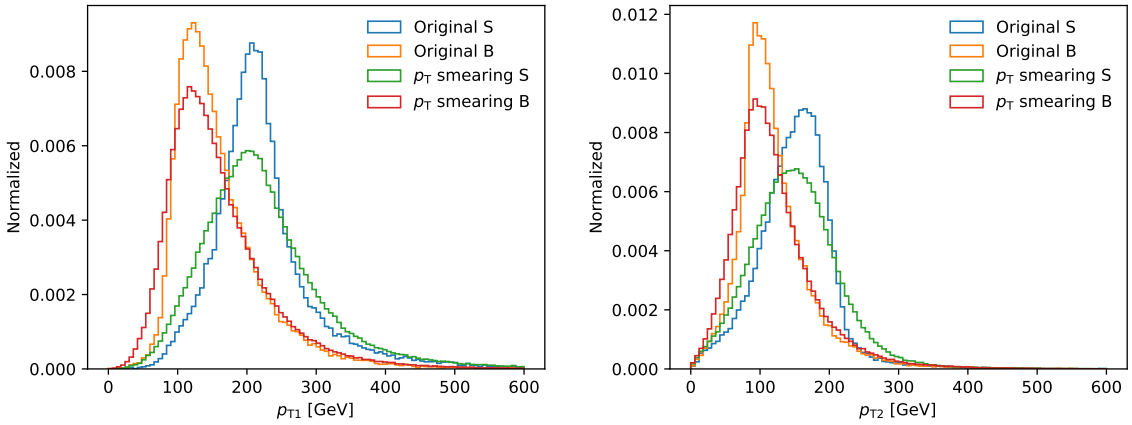
Figure 12: The pseudorapidity distribution before and after the $\eta - \phi$ smearing augmentation. η_1 and η_2 are the pseudorapidities of the leading and the sub-leading Higgs candidate, respectively.

For each type of augmentation, we test "n times augmentation" with different n. The n times augmentation means for one original sample, we generate n augmented samples. Additionally, we test another case that applies all augmentations at the same time.



(a) Leading Higgs (b) Sub-leading Higgs

Figure 13: The azimuthal angle distribution before and after the $\eta - \phi$ smearing augmentation. ϕ_1 and ϕ_2 are the azimuthal angles of the leading and the sub-leading Higgs candidate, respectively.



(a) Leading Higgs (b) Sub-leading Higgs

Figure 14: The transverse momentum distribution before and after the p_T smearing augmentation. p_{T1} and p_{T2} are the transverse momentum of the leading and the sub-leading Higgs candidates, respectively.

3.3 Training results

Table 10 presents the DNN classification training results of the original sample. Table 11 are the training results of the augmented samples. For each type of augmentation, they all can improve the ACC by about 4%. The differences among the various augmentation are not significant. The 10-times augmentation has the best results, but the difference between the 5-times and 10-times augmentation is tiny. It seems that the performance of this classifier is saturated.

Table 10: The training results of original samples. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

	Original
ACC	0.845 ± 0.015
AUC	0.917 ± 0.005

Table 11: The training results of augmented samples. "+ 3 times" means the training sample consists of the original sample and 3 times the augmented sample. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

Sample		Rotation	$\eta - \phi$ smear	p_{T} smear	All
+ 3 times	ACC	0.880 ± 0.007	0.879 ± 0.010	0.882 ± 0.003	0.875 ± 0.011
	AUC	0.950 ± 0.007	0.949 ± 0.008	0.951 ± 0.003	0.942 ± 0.012
+ 5 times	ACC	0.887 ± 0.002	0.887 ± 0.001	0.890 ± 0.002	0.889 ± 0.003
	AUC	0.955 ± 0.001	0.955 ± 0.001	0.957 ± 0.001	0.956 ± 0.001
+ 10 times	ACC	0.889 ± 0.001	0.889 ± 0.002	0.892 ± 0.002	0.892 ± 0.002
	AUC	0.956 ± 0.001	0.956 ± 0.001	0.958 ± 0.001	0.958 ± 0.000

3.4 Deeper model

In Section 3.3, the DNN model consists of 2 hidden layers, each containing 64 hidden nodes. To explore the impact of the model structure, the deeper DNN model is trained. We investigate the performance of the DNN model with 5 hidden layers.

Table 12 are the training results with a deeper DNN model. Models are only trained on the "All augmentation" sample because from Table 11 we found that four augmentation methods yielded similar results. The results show that the augmented sample can improve

Table 12: The training results of deeper DNN model. "+ 3 times" means the training sample consists of the original sample and 3 times the augmented sample. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

Sample		Original	+ 3 times	+ 5 times	+ 10 times
All augmentation	ACC	0.864 ± 0.005	0.890 ± 0.002	0.890 ± 0.002	0.884 ± 0.005
	AUC	0.928 ± 0.005	0.957 ± 0.001	0.957 ± 0.001	0.949 ± 0.005

ACC to 89%, even from the "+ 3 times" case and this accuracy value is similar to the previous test. These findings suggest that the classifier may have reached a saturation point and point out the difficulty of further improving accuracy on this test sample.

4 Hidden Valley model

4.1 Sample generation

The signal process is $f\overline{f} \to Z_V$, where Z_V is the massive gauge boson linking SM and the dark sector. The hidden Z_V boson would decay to a pair of dark quark $q_V\overline{q}_V$, leading to two jets in the detector. The signal sample is generated by `Pythia` and the detector simulation is done by `Delphes`. The anti-k_t algorithm is utilized for jet reconstruction with parameter $R = 0.8$. Some parameters are listed in Table 13.

Table 13: The parameter setting for the Hidden Valley model. "490010x:m0 = 10.3306; x=1,2,3" means x should be replaced by 1,2,3 in Pythia card.

Parameter	Value	Pythia card
M_{Z_V}	5.5 TeV	4900023:m0 = 5500
$\sqrt{s}$	13 TeV	
Λ_{D}	10 GeV	HiddenValley:Lambda = 10.0
$m_{\pi_{\mathrm{D}}}$	10 GeV	4900x1:m0 = 10.0; x=11,21,31,22,32,33
$m_{\rho_{\mathrm{D}}}$	26.944 GeV	4900x3:m0 = 26.944; x=11,21,31,22,32,33
$m_{q,\mathrm{constituent}}$	10.3306 GeV	490010x:m0 = 10.3306; x=1,2,3

The background sample is the SM QCD di-jet. This process is generated at $\sqrt{s} =$ 13 TeV. Following are the `MadGraph` scripts for generating background samples:

```
generate p p > j j
```

```
output ppjj
launch ppjj

shower=Pythia8
detector=Delphes
analysis=OFF
madspin=OFF
done

Cards/delphes_card_CMS.dat

set run_card nevents 10000
set run_card ebeam1 6500.0
set run_card ebeam2 6500.0

set run_card ptj 700
set run_card etaj 2.2
set run_card mmjj 3000

done
```

4.2 Problem for generating signal sample

Error messages:

```
PYTHIA Error: input string not found in settings databases::
  HiddenValley:separateFlav    = on! Consider different flavours

PYTHIA Error: input particle not found in Particle Data Table:
  4900102:m0                   = 10.3306

...
```

Solution: This problem arises from the `Pythia` version. At first, `Pythia 8.306` is used to generate signal samples. Some features are not included in this version. We should use `Pythia 8.307` at least. More details between 8.306 and 8.307 can be found in this page.

4.3 Sample selection

The selection cuts after the `Delphes` simulation:

- n_j cut: The number of jets should be greater than or equal to 2.
- p_{T} cut: The transverse momentum of two highest p_{T} jets should greater 750 GeV.
- η cut: The η range of two highest p_{T} jets are require $|\eta| < 2$.
- Signal region: Total invariant mass of two leading jets m_{jj} belonging to $[4700, 5500]$.
- Sideband region: Total invariant mass of two leading jets m_{jj} belonging to $[4300, 4700] \cup [5500, 5900]$.

Table 14 is the cutflow number at different selection cuts. Figure 15 is transverse momentum distribution after n_j cut. Figure 16 is the m_{jj} distribution after the η cut.

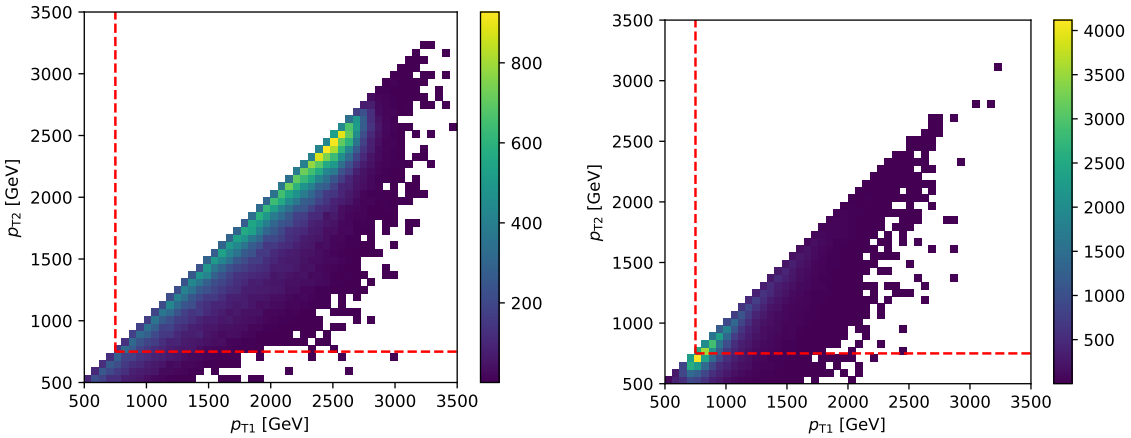
Table 14: The number of passing events and passing rates for signal and background processes at different selection cuts.

Cut	Signal	pass rate	Background	pass rate
Total	100000	1	100000	1
n_j cut	99996	1.00	99963	1.00
p_{T} cut	90901	0.91	57832	0.58
η cut	89800	0.90	55523	0.56
SR region	55844	0.56	1991	0.02
SB region	16079	0.16	3090	0.03

4.4 Jet image

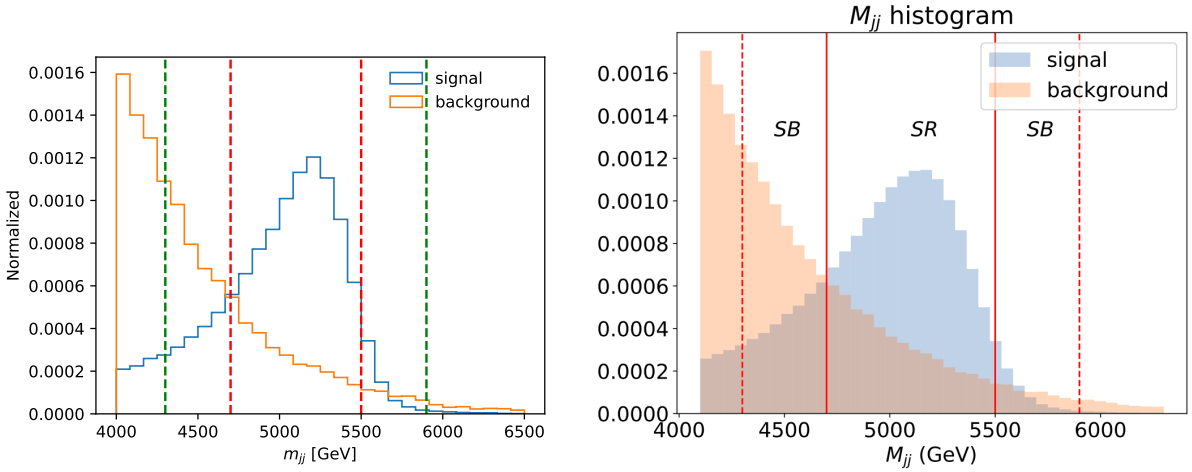
We can construct the jet image from the event passing the selection cuts described in Section 4.3. The jet image is constructed for each jet separately so that we would obtain two for each event. They are treated as two channels of a picture. The following steps construct the jet image:

1. Centralization: Compute the p_{T} weighted center in (η, ϕ) coordinates, then shift this point to origin.
2. Rotation: Rotate the highest intensity axis to the η axis.



(a) Signal (b) Background

Figure 15: The transverse momentum distribution of leading and sub-leading jets. The red dashed lines are the p_{T} cut.



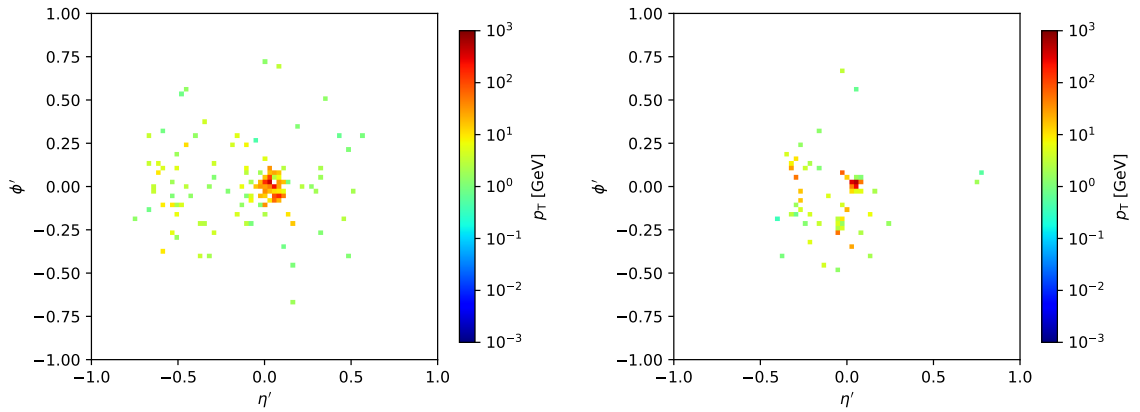
(a) My plot (b) Zong-En's plot

Figure 16: The total invriant mass m_{jj} distribution of signal and background samples. In my plot, the signal region is between the red dashed lines. The sideband region is between the green dashed lines and excludes the signal region.

3. Flipping: Flip the highest intensity quadrant to the first quadrant.

4. Pixelating in $\eta \in [-1, 1],\ \phi \in [-1, 1]$ box, with 75×75 pixels.

Figure 17 is the jet images of a single event. Figure 18 is the average jet image.



(a) Signal

(b) Background

Figure 17: The jet images of the leading jet in the signal region. The η' and ϕ' are the coordinates after the preprocessing (centralization, rotation, flipping).

4.5 Datasets

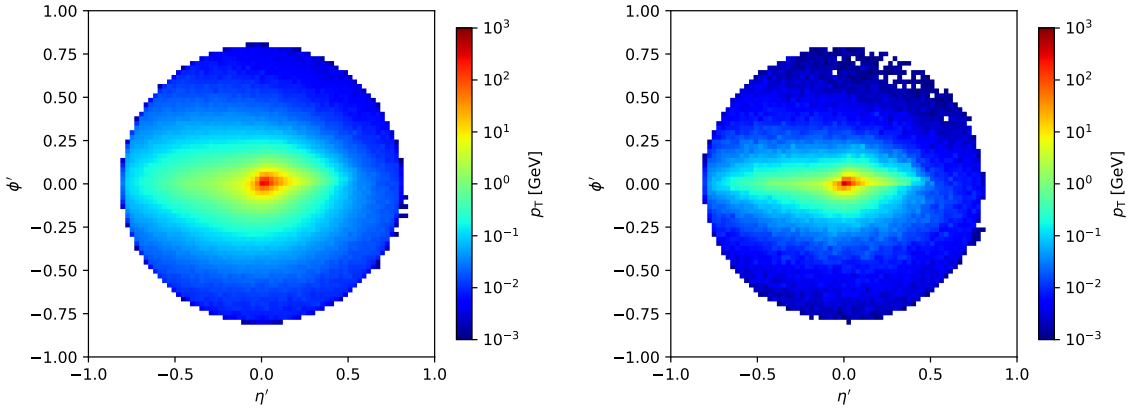
The total cross-section of the background events is 6837 fb. From Table 14, we can compute the corresponding cross-sections of signal and sideband regions are 136.1 fb and 211.2 fb, respectively. Thus, the numbers of events in signal and sideband region are 19k and 29k at luminosity $\mathcal{L} = 139$ fb^{-1}.

The training sample sizes are summarized in Table 15.

Table 15: The training sample size for the mixed sample. We set sensitivity $S/\sqrt{B} = 1$ in the signal region and evaluate the number of events in the signal region. Then, the number of events in the sideband region can be obtained from Table 14.

	True label	
Mixed sample	Signal	Background
Signal region	138	19k
Sideband region	40	29k

The pure testing sample consists of 10k signal events and 10k background events selected from the signal region.



(a) Signal

(b) Background

Figure 18: The average jet images of the leading jet in the signal region. The η' and ϕ' are the coordinates after the preprocessing (centralization, rotation, flipping). The number of the signal events is 56k. The number of the background events is 20k.

4.6 Training CNN

The CNN model structure is summarized in Figure 19. The internal node uses the rectified leaner unit (ReLU) as the activation function. The loss function is Categorical cross-entropy, and we use the Adam optimizer to optimize the loss function.

4.7 Hidden Valley model training results

The CNN is trained on samples with sensitivity $S/\sqrt{B}$ ranging from 1 to 10. Figure 20 presents the training results. These numbers are evaluated from the pure samples. The CNN cannot learn useful information for the case with a sensitivity $S/\sqrt{B} \leq 5$, so the ACC is 0.5. After this region, the accuracy demonstrates a significant improvement. It seems that the CNN model surpasses a certain threshold.

4.8 Data process procedure

1. Generate the sample file in `.root` format. Following Section 4.1.

2. Apply the selection cuts described in Section 4.3 and save the event passing the cuts in `HDF5` format. The file contains the information listed below

   - The (p_T, η, ϕ) of leading and sub-leading jet constituents.
   - Total invariant mass m_{jj}.

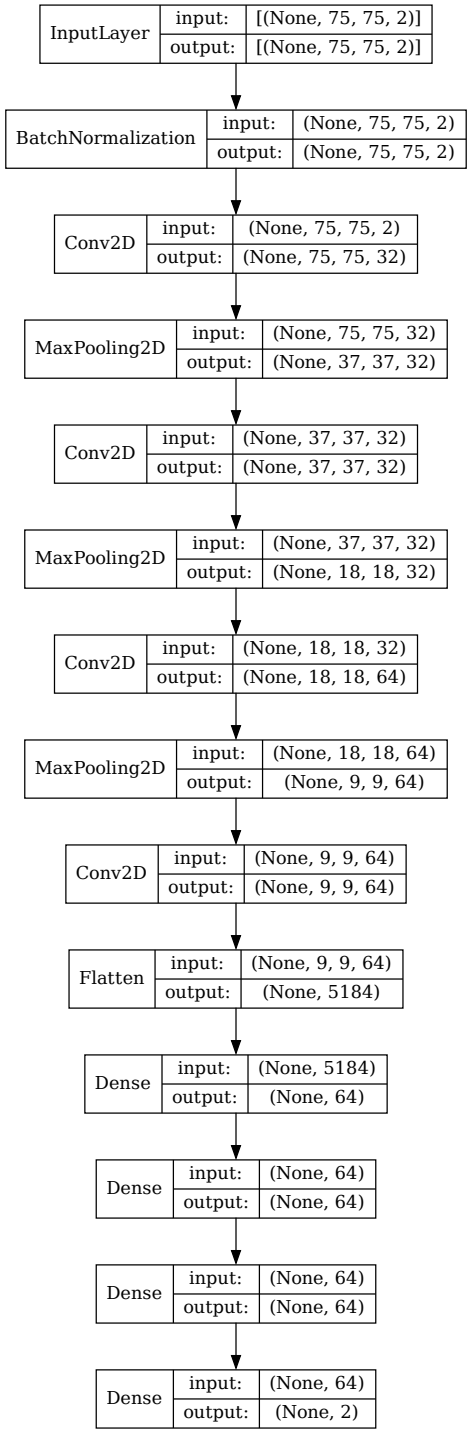


Figure 19: For first and second Conv2D layers, the kernel sizes are 5×5. For third and fourth Conv2D layers, the kernel sizes are 3×3.

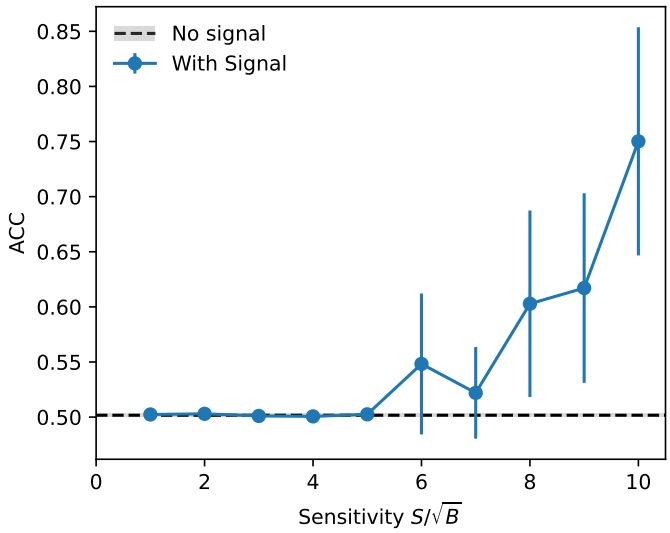


Figure 20: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. For sensitivity less than 5, the error bar is too small to see.

- Type of event: 1 for signal, 0 for background.

3. Make mixed sample in `HDF5` format. Following Section 4.5, we can compute the size of datasets.

4. Apply data augmentation in `HDF5` format. Following Section 4.9.

5. Generate the jet image from `HDF5` data.

Note the preprocessing is done in step 5. If we do the preprocessing many times, the rounding errors would break the jet image.

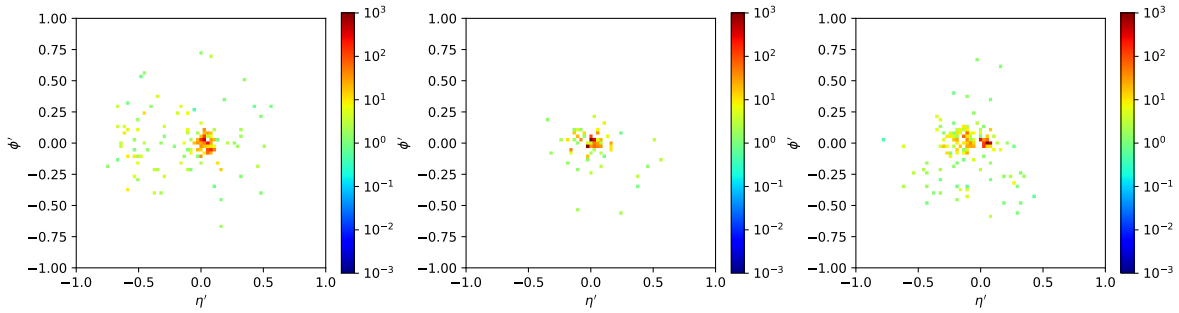
4.9 Data augmentation

To reduce the threshold, the data augmentation technique is tested. Similar to the Section 3.2, we apply the $\eta - \phi$ smearing on our training sample. Specifically, the (η, ϕ) coordinate of jet constituents are resampled according to a Normal distribution centered on the original coordinate and with a standard deviation inversely proportional to the p_{T}

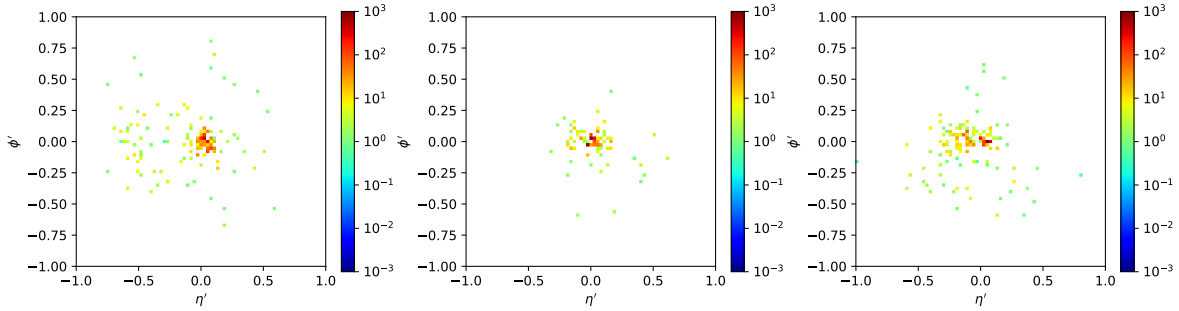
$$\eta' \sim \mathcal{N}\left(\eta, \frac{\Lambda}{p_{\mathrm{T}}}\right), \quad \phi' \sim \mathcal{N}\left(\phi, \frac{\Lambda}{p_{\mathrm{T}}}\right) \tag{4}$$

where η', ϕ' are the augmented coordinate, p_{T} is the transverse momentum of the jet constituent, and the smearing scale is set to be $\Lambda = 100$ MeV.

Figure 21 and 22 are the jet image before and after the augmentation. The jet images look similar before and after the augmentation, but not the same.



(a) Before augmentation



(b) After augmentation

Figure 21: The jet images before and after the $\eta - \phi$ smearing augmentation.

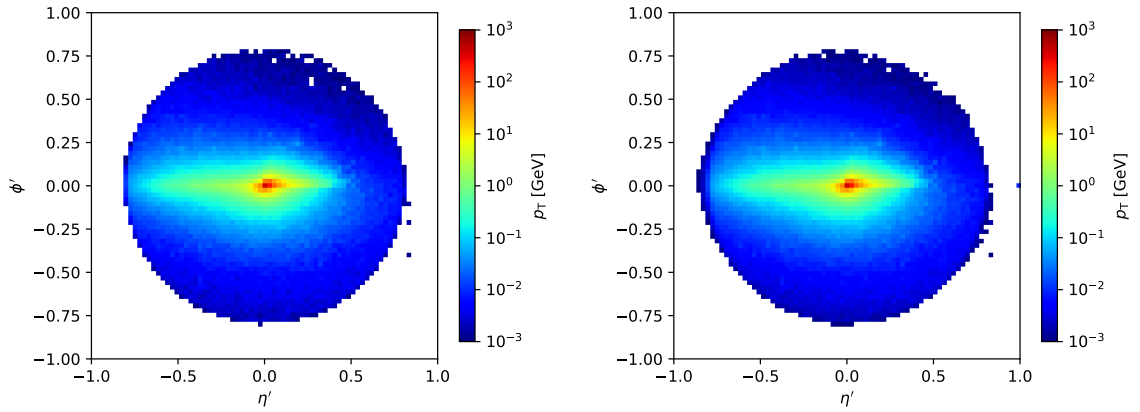
We generate samples with sensitivity $S/\sqrt{B}$ ranging from 1 to 10. Then, apply the data augmentation to make larger samples. These samples are used for CNN training. Figure 23 presents the training results. These numbers are evaluated from the pure samples. The ACC is better than previous results (blue curve in Figure 20) for the "+1 times" curve for all sensitivities. The data augmentation technique suppresses the threshold and improves training performance.

Figure 24 presents the training results with larger samples. For the "+3 times" curve, there is a strange result, the ACC is worse than the "+1 times" case for sensitivities $S/\sqrt{B} \geq 3$. Here are some things that could be improved in the training procedure, e.g. overfitting.

Figure 25 shows the average jet image for different augmentation times.

4.10 Enlarge the training sample size

To validate the correctness of the data augmentation results, we generated larger samples by scaling the luminosity. To compare the results with "+1 times" augmentation, we scale



(a) Before augmentation

(b) After augmentation

Figure 22: The average jet images before and after the $\eta - \phi$ smearing augmentation.

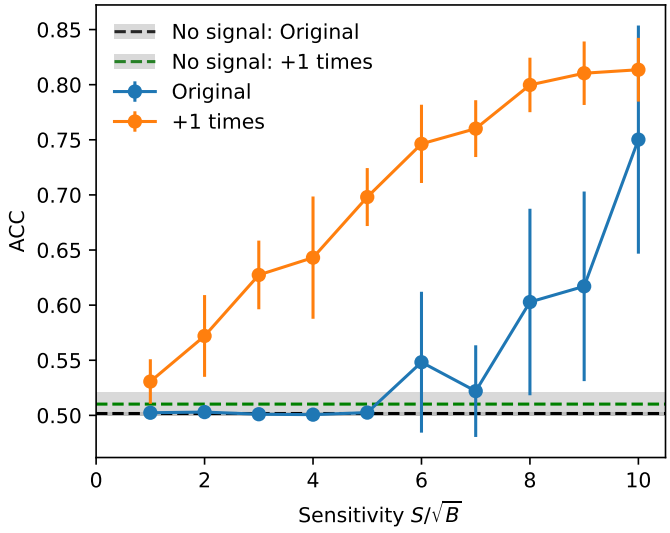


Figure 23: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

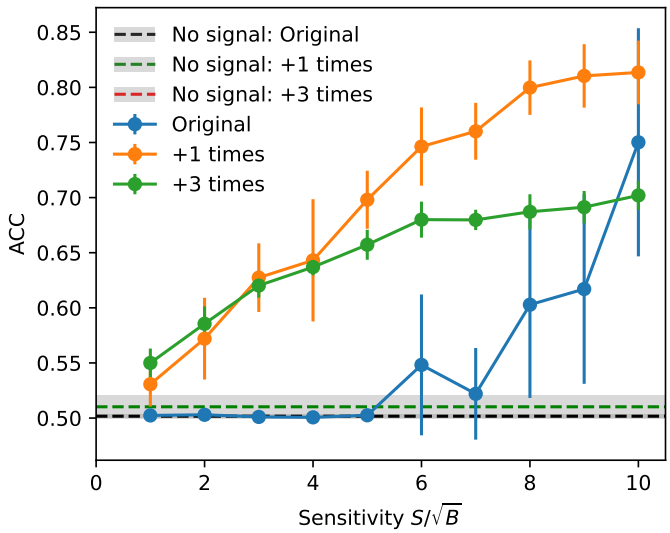
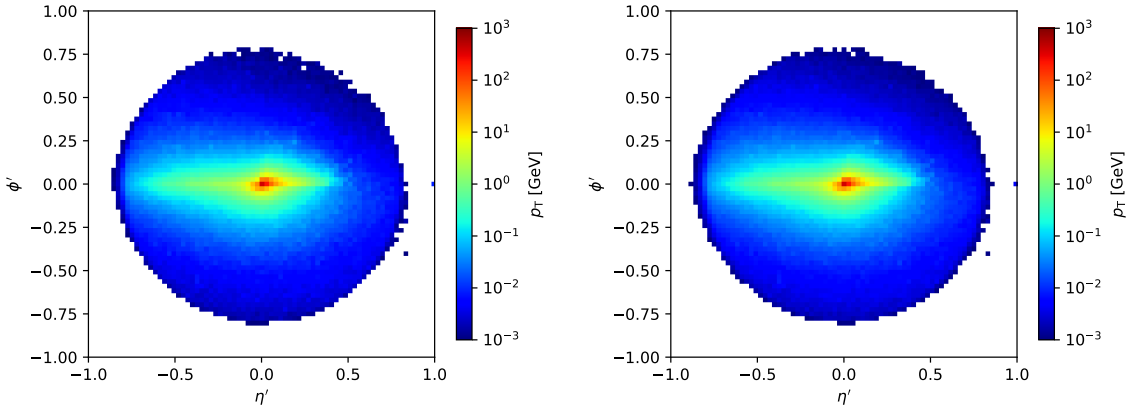


Figure 24: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.



(a) +1 times (b) +3 times

Figure 25: The average jet images of the $\eta - \phi$ smearing augmentation with different times.

the luminosity to $\mathcal{L} = 139 \times 2$ fb^{-1}. It is important to note that the number of signal events is not equal for "+1 times" and "luminosity ×2" at the same sensitivity because the number of signal events is computed from $S/\sqrt{B}$ for a given B, resulting in the number of "+1 times" signal sample is greater by a factor of $\sqrt{2}$. Table 16 is an example. For a fair comparison, where the sizes of signal and background events are the same, we should compare the "+1 times" results with the point on the "luminosity ×2" curve corresponding to $\sqrt{2}$ times the sensitivity.

Table 16: The training sample size for the original, "+1 times" and "luminosity ×2" mixed sample. We set sensitivity $S/\sqrt{B} = 1$ in the signal region and evaluate the number of events in the signal region.

	Original		+1 times		luminosity ×2	
Mixed sample	Signal	Background	Signal	Background	Signal	Background
Signal region	138	19k	276	38k	194	38k
Sideband region	40	29k	80	58k	56	58k

Figure 26 presents the training results with larger samples. The results of "luminosity ×2" is better than the original sample, but still worse than the "+1 times" case. Even considering the $\sqrt{2}$ scale factor, the corresponding points' performance is still worse. There may be something wrong with the data processing or training codes.

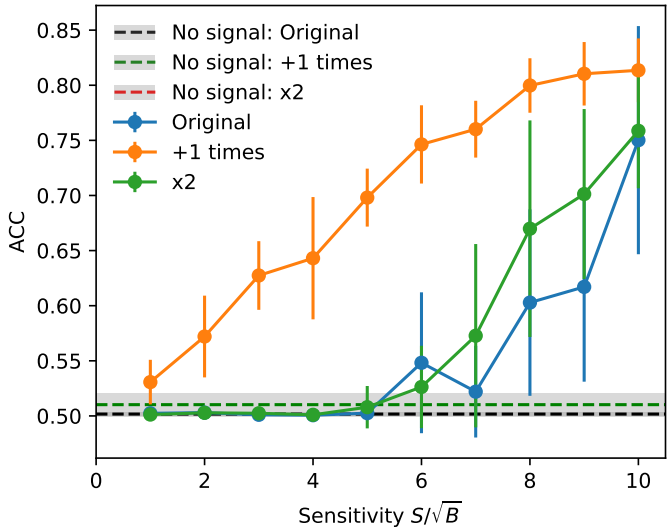


Figure 26: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

4.11 Check the codes and make more plots

The data process steps from 2 to 5 (Section 4.8) have been checked and samples are processed again.

Figure 27 is the p_{T} distribution of the jet constituents. The distributions of leading and sub-leading jet constituents are similar.

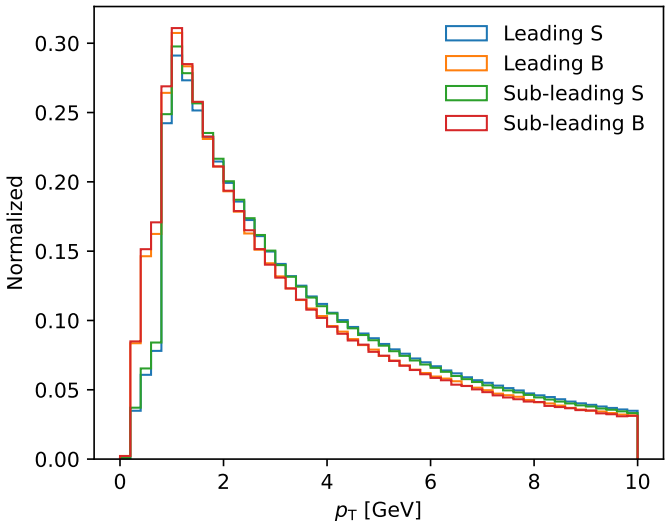


Figure 27: The transverse momentum distribution of jet constituents.

Figure 28 and 29 are the η, ϕ distributions before and after the augmentation. Because the smearing scale is small, the distributions look almost the same for both cases.

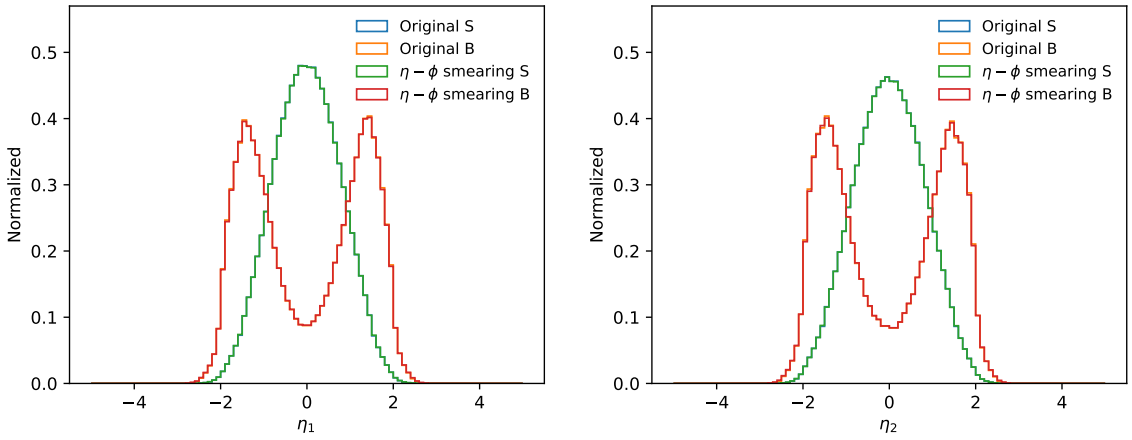
Figure 30 shows the average jet image for various cases.

Figure 31 is training results with re-processed samples.

4.12 Sensitivity difference

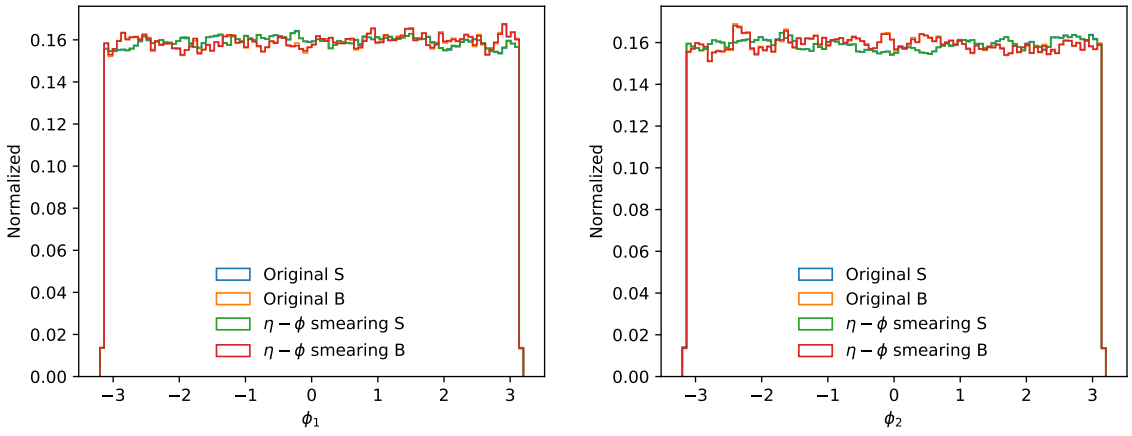
The CWoLa CNN is applied for event selection. The threshold is set for a given background efficiency ε_{b} (false positive rate). For a given ε_b, the sensitivity would be scale by a factor $\varepsilon_{\mathrm{s}}/\sqrt{\varepsilon_{\mathrm{b}}}$.

Figure 32 is the sensitivity improvement of orginal, "+1 times" and "luminosity $\times 2$" samples. The 10% curves give more stable results. The lower background efficiencies could achieve better results at the range with higher sensitivities, but the standard deviation would be much larger. Figure 33 compares the sensitivity improvement of various training samples with the same background efficiency. The "+1 times" samples provide the best threshold among all.



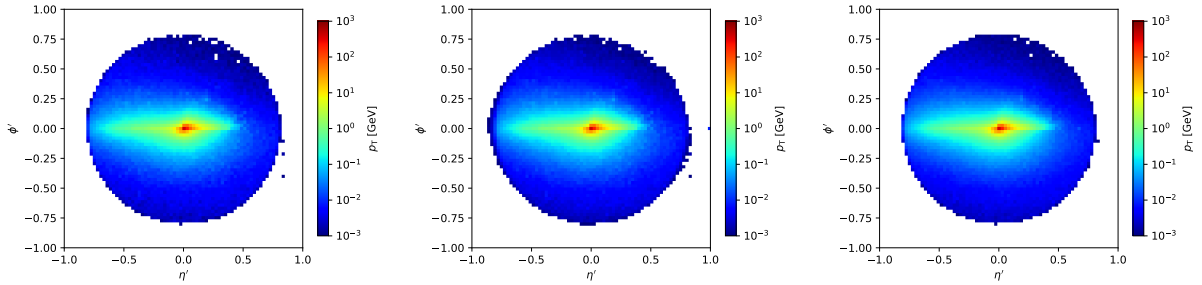
(a) Leading Higgs (b) Sub-leading Higgs

Figure 28: The pseudorapidity distribution before and after the $\eta-\phi$ smearing augmentation. η_1 and η_2 are the pseudorapidities of the leading and sub-leading jet constituents.



(a) Leading Higgs (b) Sub-leading Higgs

Figure 29: The azimuthal angle distribution before and after the $\eta-\phi$ smearing augmentation. ϕ_1 and ϕ_2 are the azimuthal angles of the leading and sub-leading jet constituents.



(a) Original (b) +1 times augmentation (c) Luminosity ×2

Figure 30: The average jet images of different cases. These plots are made from mixed samples.

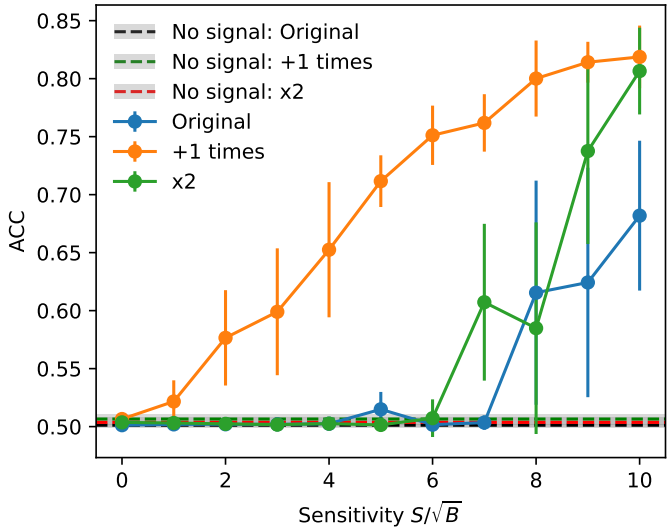
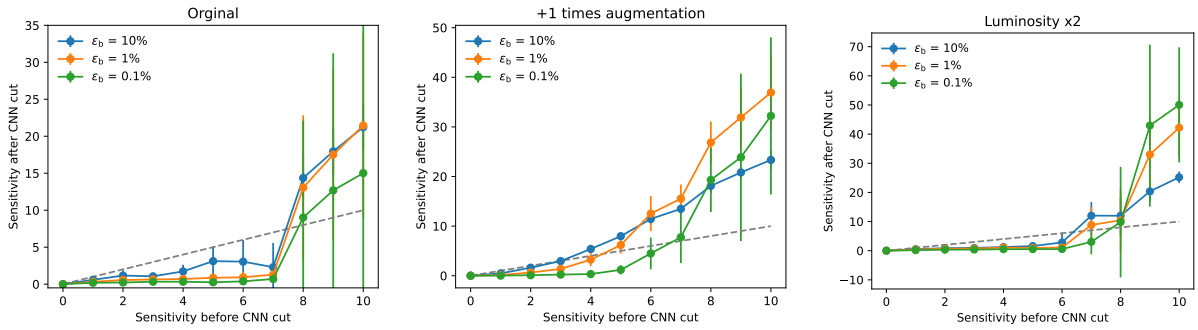


Figure 31: The performance of CWoLa CNN training with re-processed samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.



(a) Original (b) +1 times augmentation (c) Luminosity $\times 2$

Figure 32: The sensitivities before and after the CWoLa CNN selection. The thresholds are chosen from $\varepsilon_b = 10\%, 1\%, 0.1\%$. The slope of the dashed line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

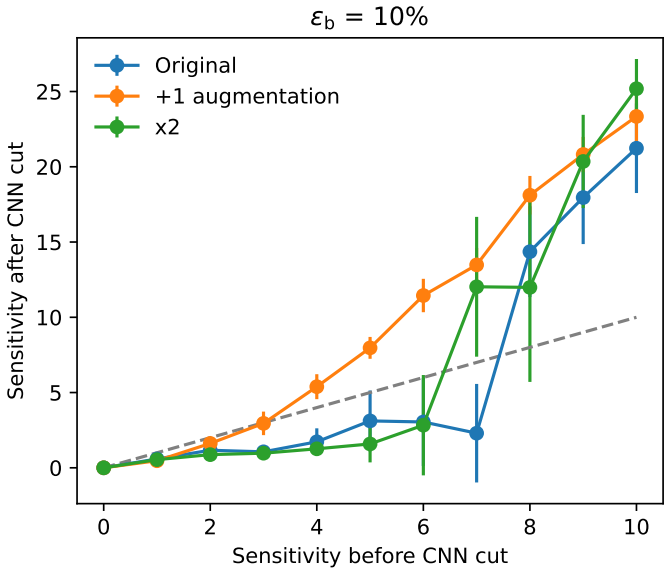


Figure 33: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "+1 times" samples provide the best threshold.

4.13 Training and true performance

In our intuition, we would expect that"+1 times" and "luminosity ×2" samples should exhibit similar performance, or the "luminosity ×2" sample might provide better results since the data augmentation generates more samples from the known data. However, the results in Figure 33 contradict our intuition. Since we use the CWoLa approach, our comparison is not based on the training performance on mixed samples; instead, we evaluate the testing performance on pure samples, referred to as "true performance." The relationship between training and true performance may not be a monotonic function and can be complicated.

Figure 34 is the scatter plot for training and true accuracy. Training ACC values are close for all cases, but true ACC values can differ greatly. The true ACC is ten times larger than the training ACC range. For training ACC, both "+1 times" and "luminosity ×2" samples show similar performance, slightly better than the original results. For the true ACC, the "+1 times" provides much better results. It seems that "+1 times" samples are more effective in improving true ACC.

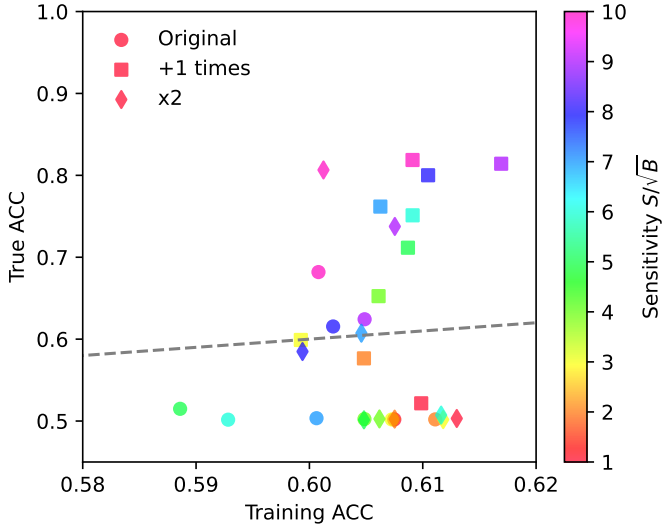


Figure 34: Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC. The range of training ACC is tiny, roughly 0.03, while the range of true ACC is about 0.3.

Figure 35 includes results with"+3 times" training . The "+3 times" samples provide much better training ACC but do not achieve higher performance for true ACC.

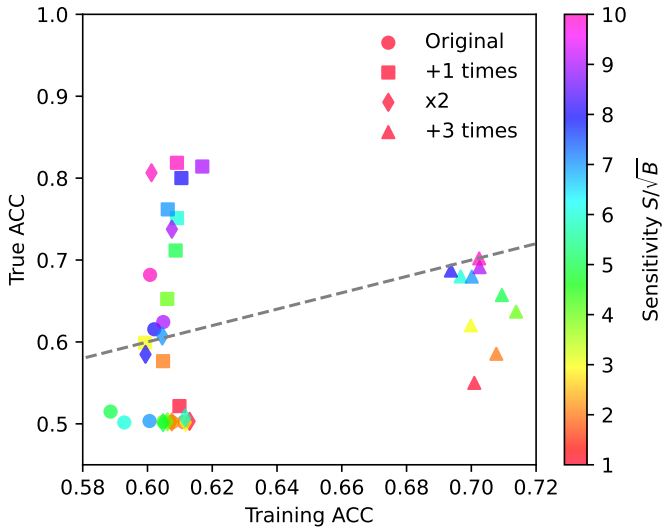


Figure 35: Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

4.14 Miscellaneous test

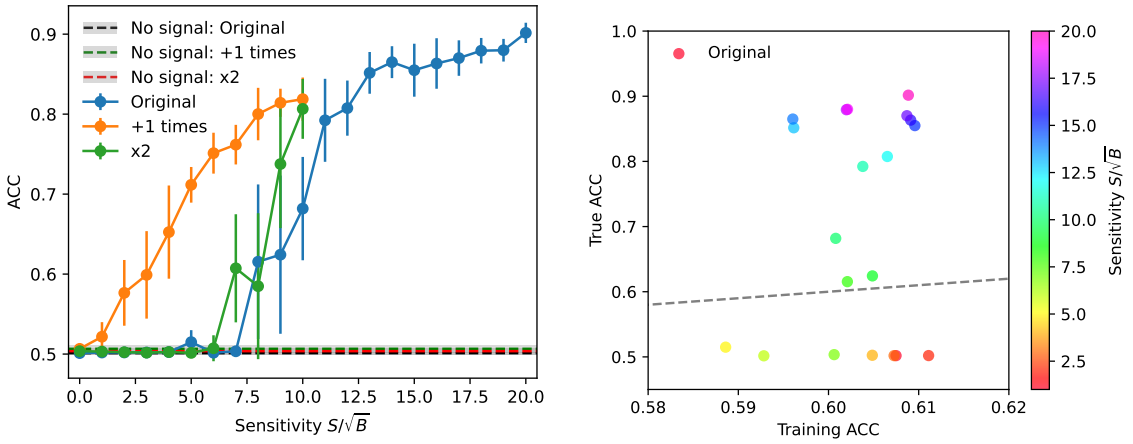
4.14.1 Sensitivity 11 to 20

To find out the training performance when sensitivity is greater than 10, we generate the samples with sensitivity ranging from 11 to 20. The training performance is summarized in Figure 36. The ACC is increased when the sensitivity is increased. When the $S/\sqrt{B} > 13$, the ACC improvement is slowed, where the ACC is 85%. The training ACC is still similar for all cases, but the true ACC exhibits a better value when the sensitivity is higher.

4.14.2 +1 copy sample

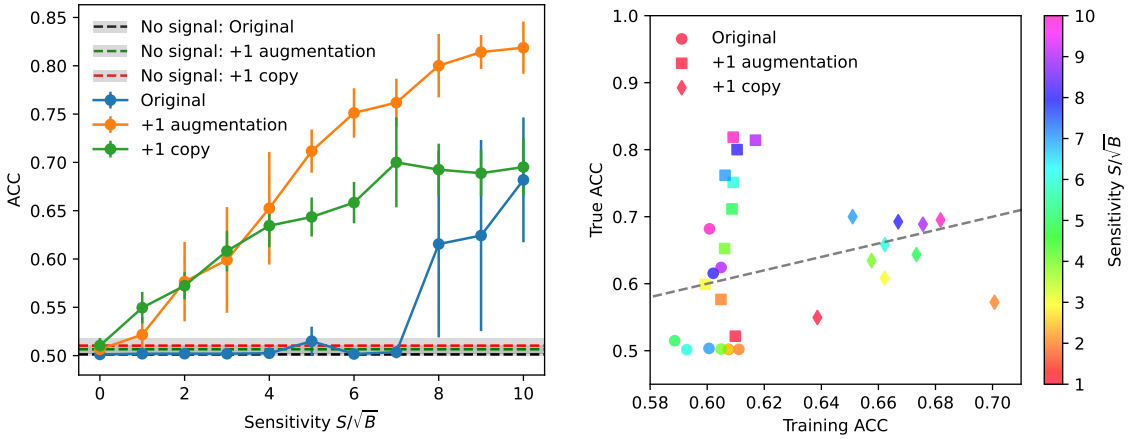
To investigate the impact of the data augmentation technique, we generate a sample by duplicating the original sample without any augmentation. This sample is called "+1 copy." Figure 37 is the ACC curve and ACC distribution. The ACC is close to the "+1 augmentation" case when $S/\sqrt{B} \leq 4$, while the ACC is worse when $S/\sqrt{B} \geq 5$. The training ACC is much better than the original and "+1 augmentation" samples, but the true ACC can not achieve higher values.

Figure 38 is the sensitivity improvement. The performance of the "+1 copy" sample is worse than "+1 augmentation". It exhibits the worst results in high-sensitivity regions. The data augmentation is effective in improving the true ACC.



(a) (b)

Figure 36: (a) The performance of CWoLa CNN training with higher sensitivity samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.



(a) (b)

Figure 37: (a) The performance of CWoLa CNN training with "+1 copy" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

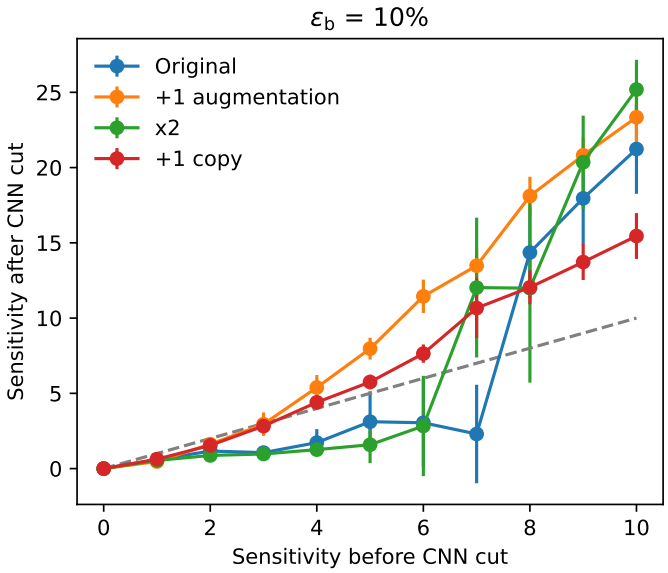


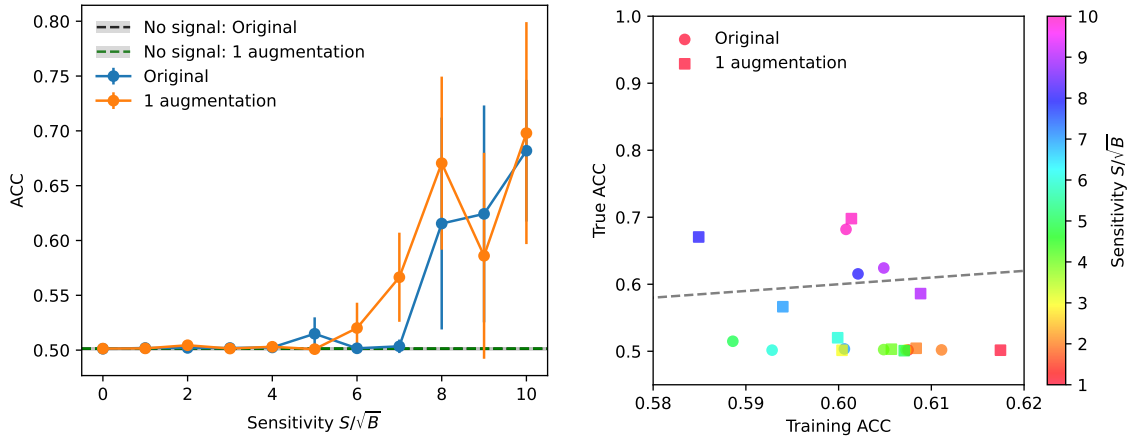
Figure 38: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "+1 times" samples provide the best threshold.

4.15 Train on only augmented sample

We train the CWoLa CNN with only augmented datasets to diagnose augmented samples. The training samples are generated from the original dataset and only the augmented part is kept for training. We test "1 augmentation" and "2 augmentation" cases. They consist of pure augmented samples. Their sizes are 1 time and 2 times the original dataset.

Figure 39 shows the ACC curve and ACC distribution with the "1 augmentation" case. The ACC of the "1 augmentation" sample is better than the "original" one at $S/\sqrt{B} = 7$, but they perform similarly on other sensitivities. Figure 40 is the sensitivity improvement. The threshold of the "1 augmentation" sample is lower than the "original" one, but they exhibit similar performance on other sensitivities. These results are consistent with the ACC curve. The difference at $S/\sqrt{B} = 7$ may come from the training fluctuation.

Figure 41 is the training results with "2 augmentation" samples. The results are similar to the "+1 augmentation" case. Figure 42 is the sensitivity improvement. The performance of the "2 augmentation" samples is almost the same as "+1 augmentation".



(a) (b)

Figure 39: (a) The performance of CWoLa CNN training with "1 augmentation" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

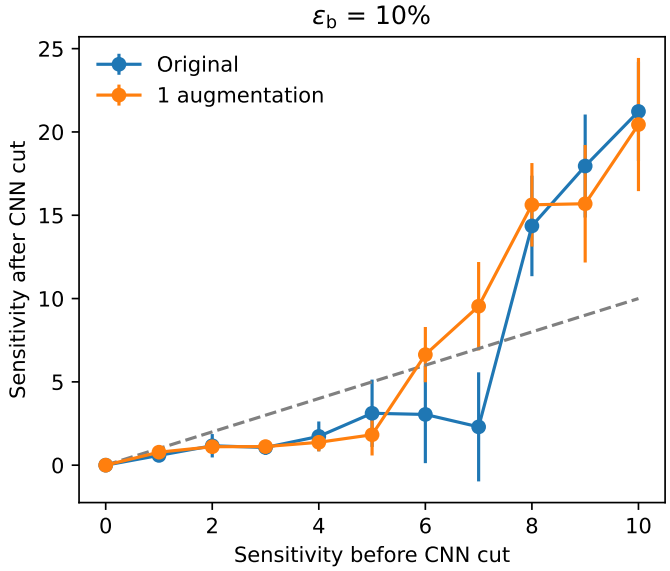
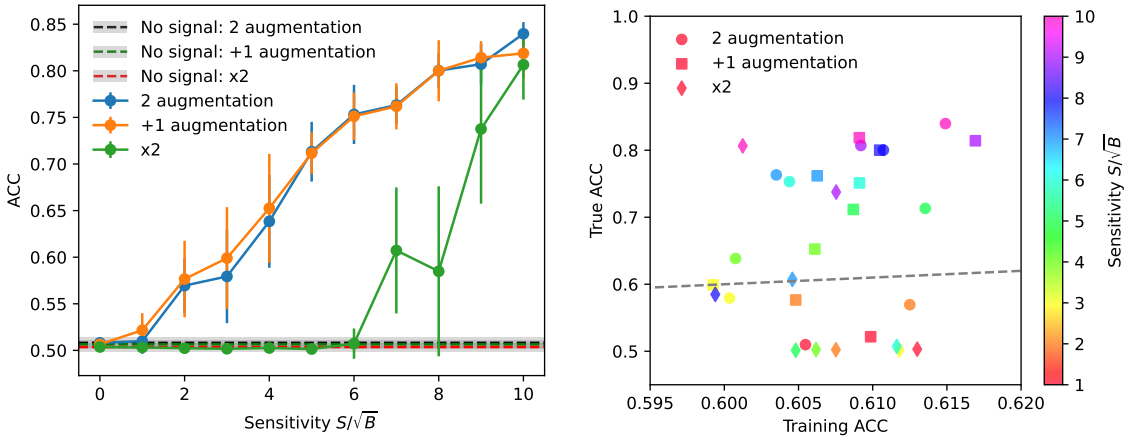


Figure 40: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "1 augmentation" samples provide a lower threshold.



(a) (b)

Figure 41: (a) The performance of CWoLa CNN training with "2 augmentation" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

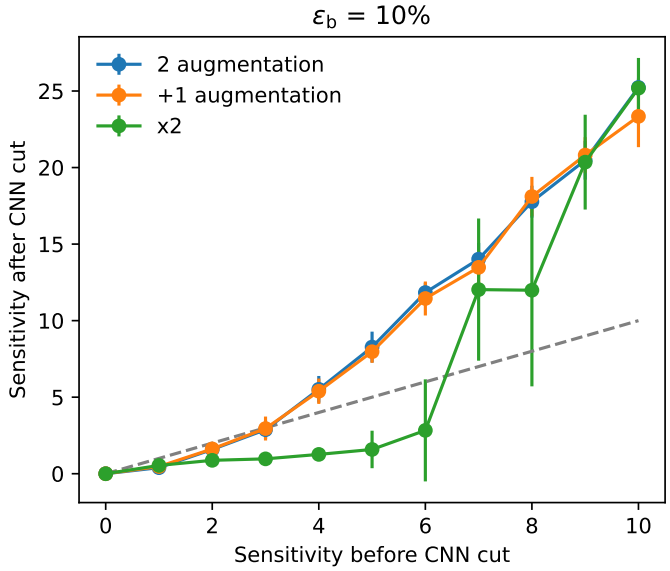


Figure 42: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "2 augmentation" and "+1 augmentation" samples perform similarly.

4.16 Change the smearing scale

To investigate the impact of the smearing scale Λ of data augmentation, we generate samples with various Λ for CWoLa CNN training. We test $\Lambda = 200, 500$ MeV samples.

Figure 43 and 44 are the jet image with different smearing scales. When the smearing scale increases the jet image is more different.

Figure 45 is the training results with various smearing scales. In the low-sensitivity region, the performance is better for the lower smearing scale, while in the high-sensitivity region, the performance is better for the higher smearing scale. Figure 46 is the sensitivity improvement. The results are consistent with the ACC curve. $\Lambda = 100$ MeV exhibits the lower threshold.

4.17 Duplicate training sample

To compare the results with Section 4.14.2, we generate a sample by duplicating the original without further shuffle. Therefore, the first n-th sample and $n+1$-th to the final sample should be the same. This dataset is called "+1 copy without shuffle." Figure 47 is the average jet image of the first n-th sample and $n+1$-th to the final sample. These two images are the same.

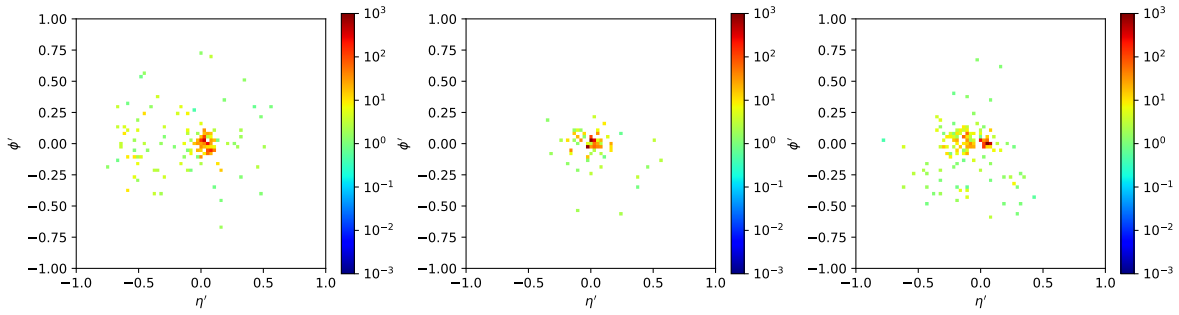
Figure 48 is the ACC curve and ACC distribution. The training results for "+1 copy" and "+1 copy without shuffle" do not have significant differences and they all perform better than the original training sample.

Figure 49 is the sensitivity improvement. The performance of the "+1 copy" and "+1 copy without shuffle" samples are similar.

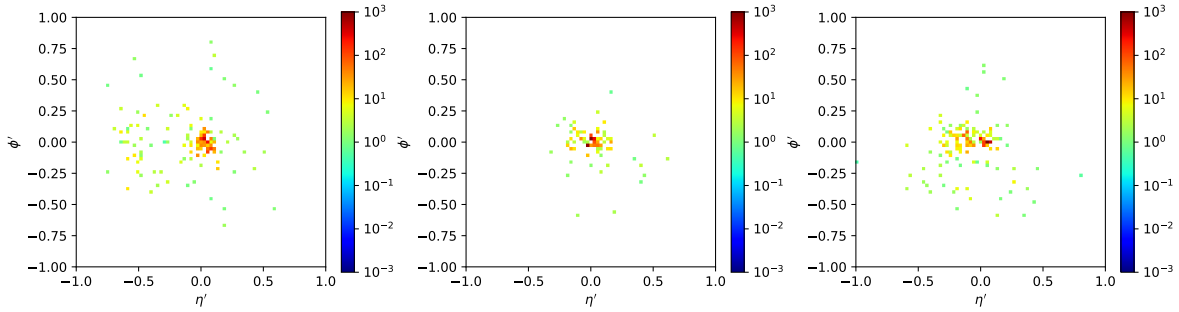
4.17.1 Remove shuffle in training code

I remove default shuffles in the training codes. In the default setting, the training sample would be shuffled before training and testing sample splitting and each epoch training. Figure 50 is the ACC curve and ACC distribution. The training results of removing shuffling codes are similar to the previous one, and all "+1 copy" samples perform better than the original.

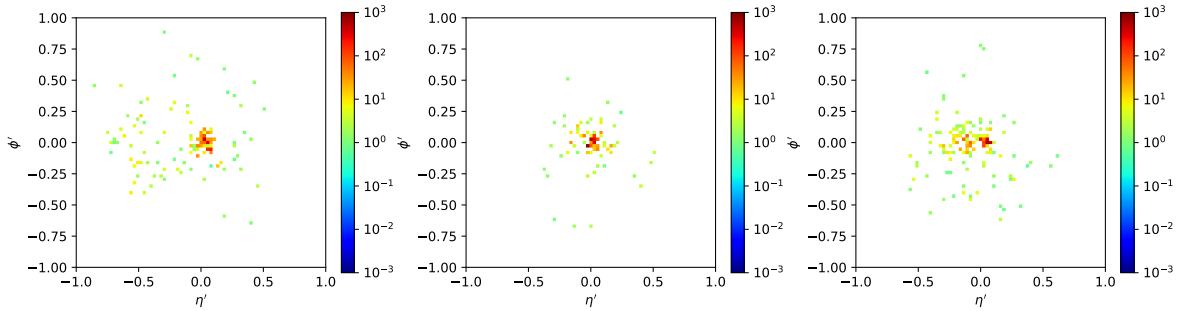
Figure 51 is the sensitivity improvement. The performance of the "+1 copy" sample and "+1 copy" with removing shuffling codes are similar.



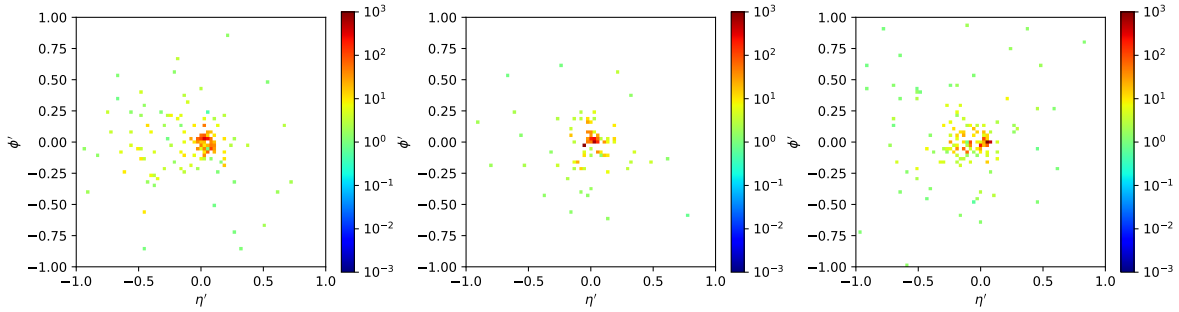
(a) Original



(b) $\Lambda = 100$ MeV

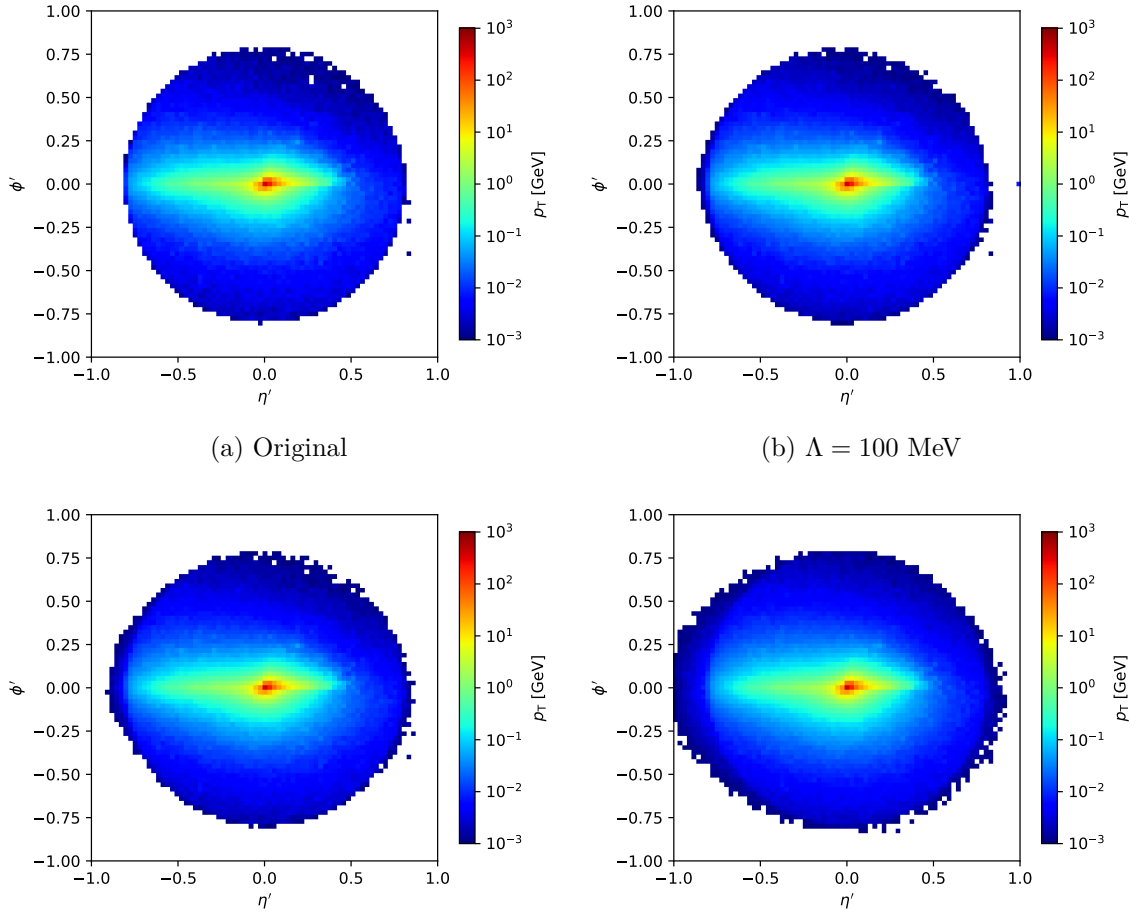


(c) $\Lambda = 200$ MeV



(d) $\Lambda = 500$ MeV

Figure 43: The jet images with different smearing scale Λ.



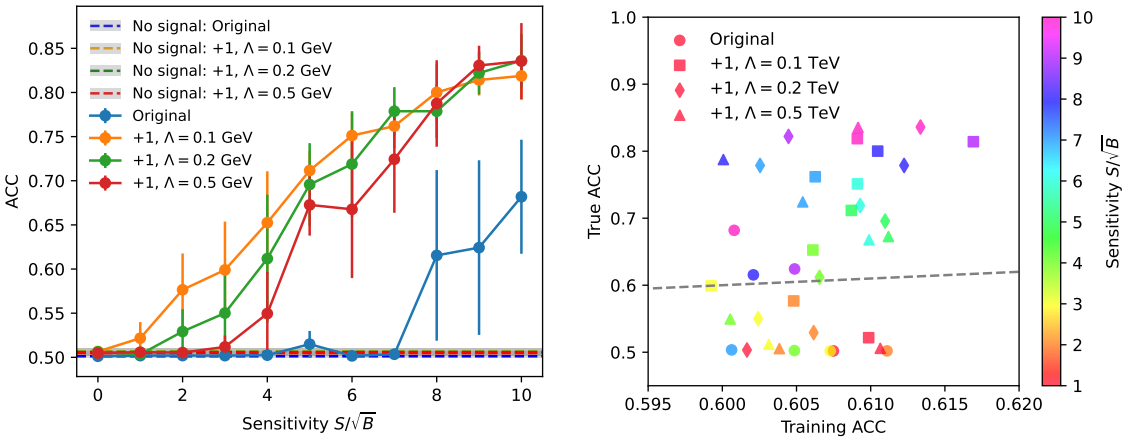
(a) Original

(b) $\Lambda = 100$ MeV

(c) $\Lambda = 200$ MeV

(d) $\Lambda = 500$ MeV

Figure 44: The average jet images with different smearing scale Λ.



(a) (b)

Figure 45: (a) The performance of CWoLa CNN training with various smearing scales. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. "+1" means the +1 time augmentation sample. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

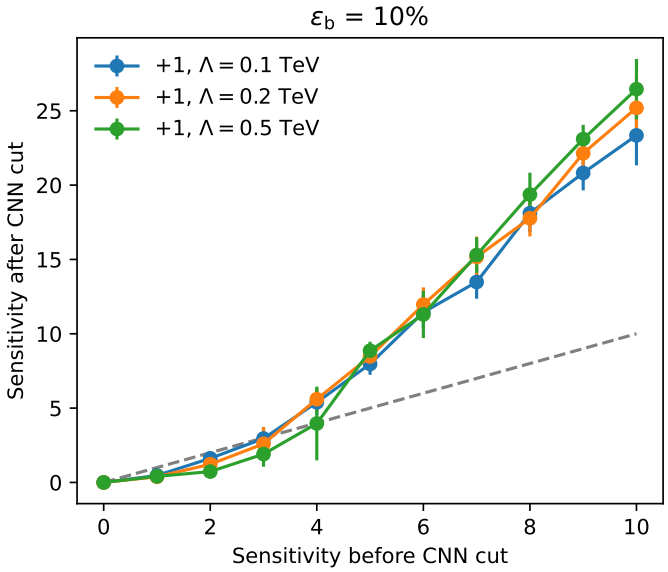
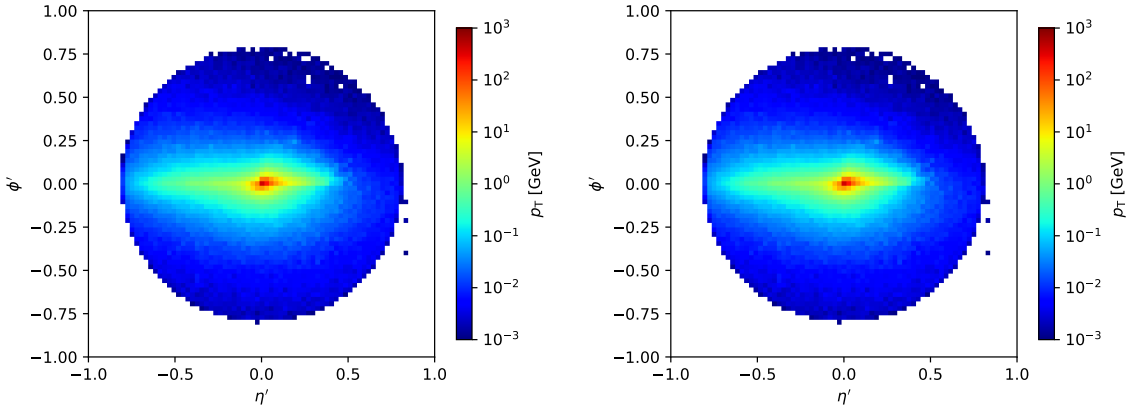
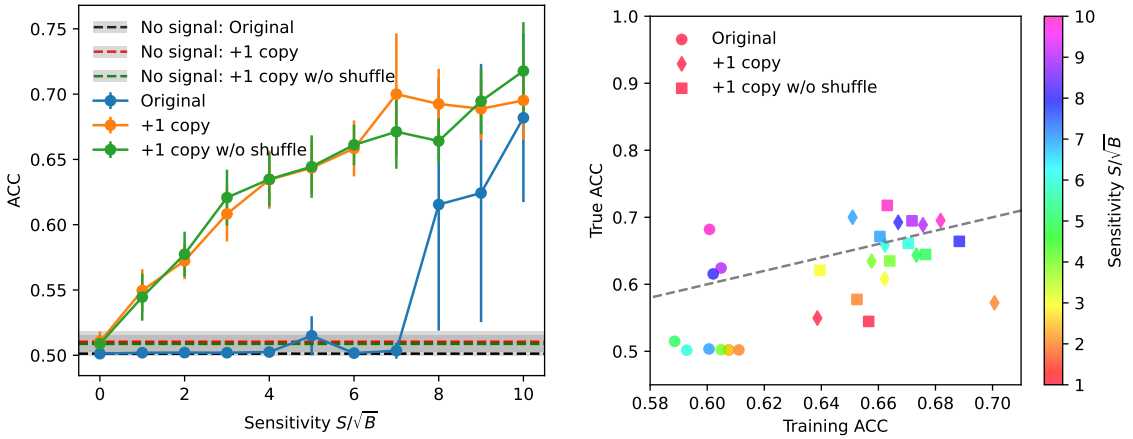


Figure 46: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. "+1" means the +1 time augmentation sample.



(a) $1 \sim n$ (b) $n+1 \sim 2n$

Figure 47: The average jet image of first n-th sample and $n+1$-th to the final sample.



(a) (b)

Figure 48: (a) The performance of CWoLa CNN training with "+1 copy" and "+1 copy without shuffle" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

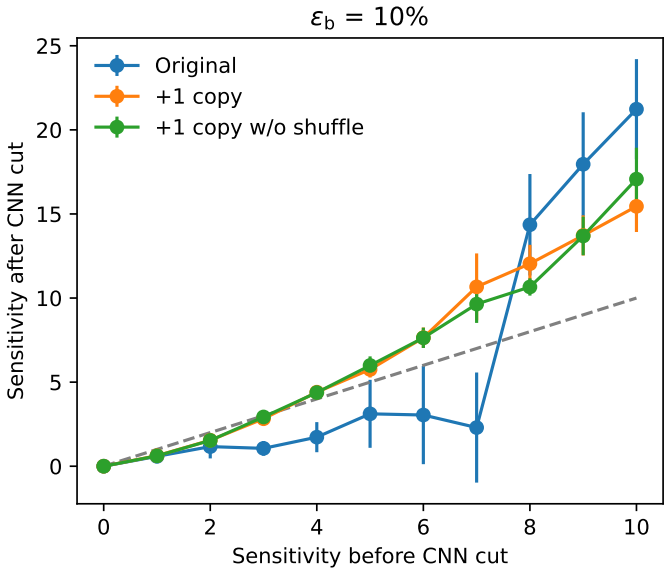
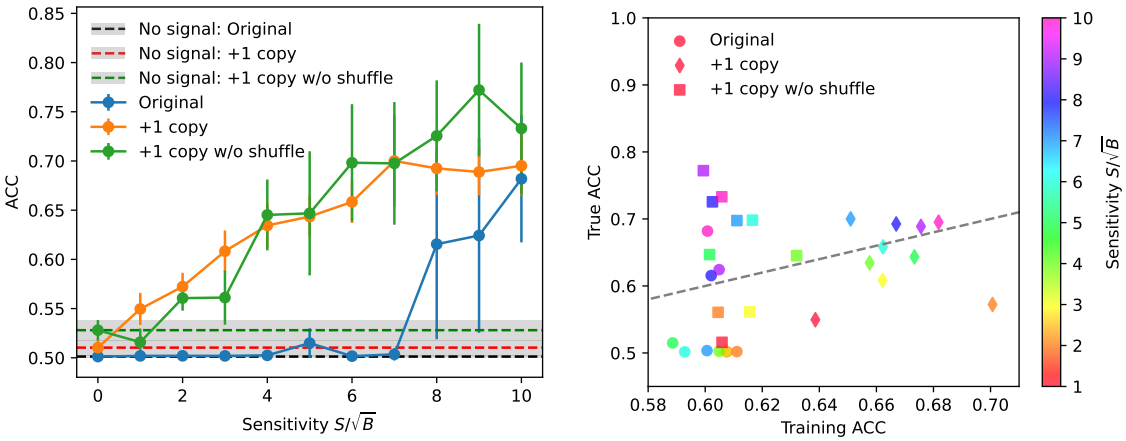


Figure 49: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.



(a) (b)

Figure 50: (a) The performance of CWoLa CNN training with "+1 copy" sample and "+1 copy" with removing shuffling codes. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

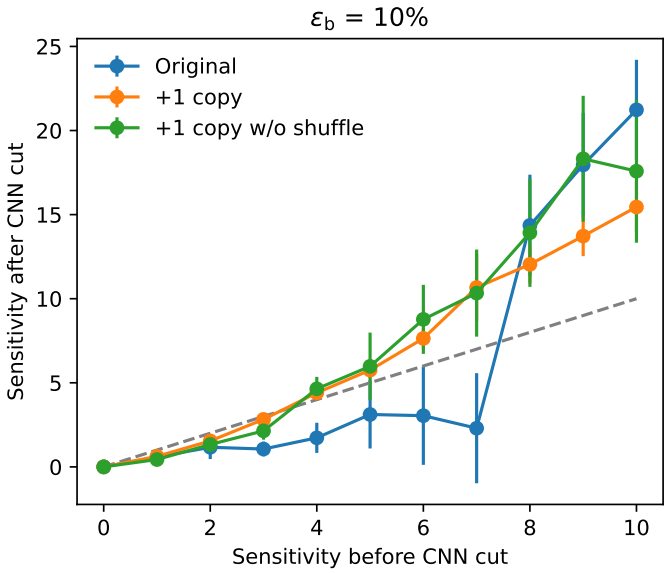


Figure 51: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.17.2 Half original + copy

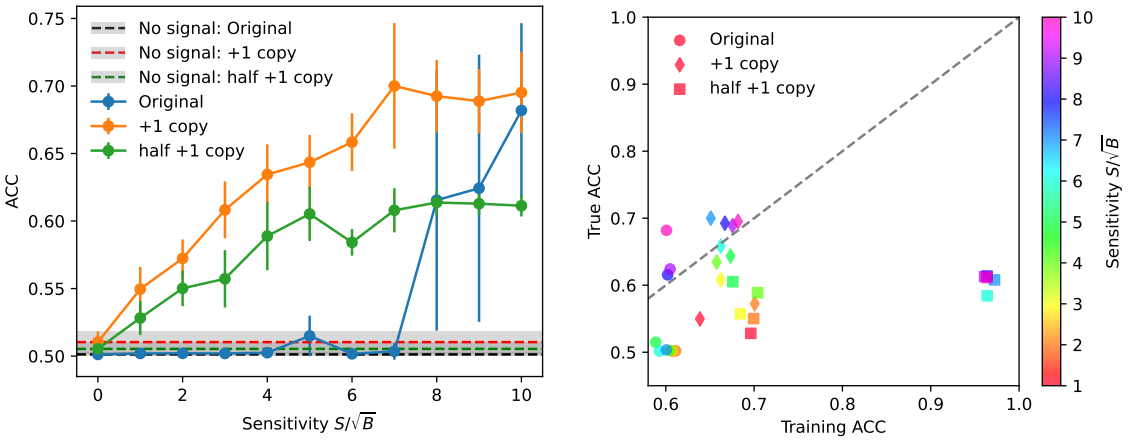
To make sure the better training results come from adding copy samples and not from the sample size, we use half the original sample and duplicate it. This dataset is called " half +1 copy" and has the same size as the original sample.

Figure 52 is the ACC curve and ACC distribution. The training results of "half +1 copy" are better than the original training sample in the low sensitivity region, while in the high sensitivity region, "half +1 copy" has the worse results. For the "half +1 copy" case, the training ACC is overfitting when $S/\sqrt{B} \geq 6$.

Figure 53 is the sensitivity improvement. The threshold of the "half +1 copy" sample is lower than the original training sample, but the sensitivity improvement is smaller than the original dataset in the high sensitivity region.

4.18 Another testing dataset

To rule out the problem in testing samples, I generate more signal and background samples and prepare another testing dataset. The testing results of old and new testing samples are shown in Figure 54. All cases perform similarly on two testing datasets. Note that because we only save the best model, we only present the ACC with the best training



(a) (b)

Figure 52: (a) The performance of CWoLa CNN training with "+1 copy" sample and "half +1 copy" datasets. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

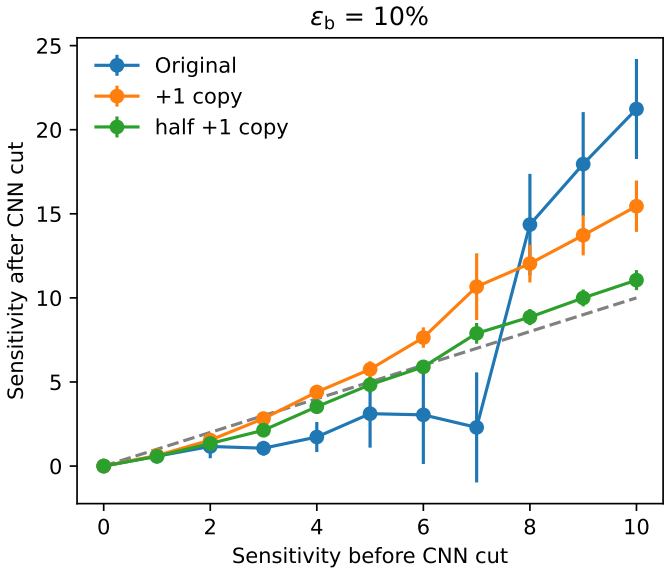
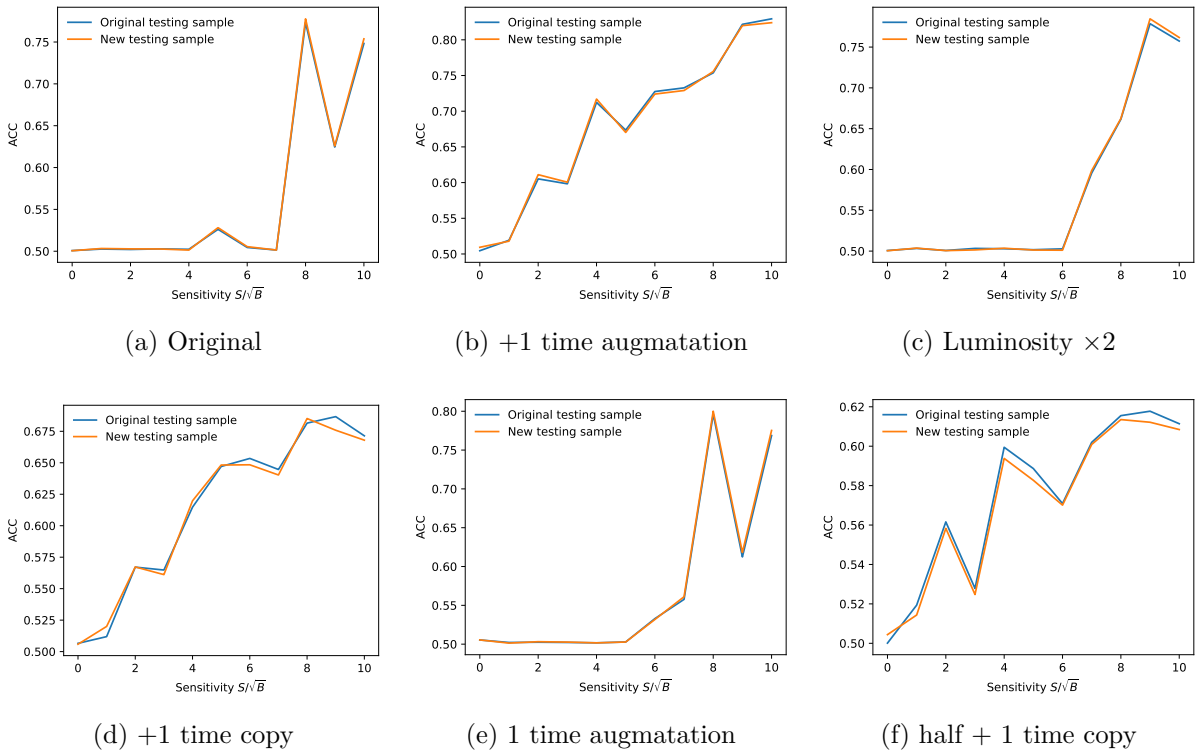


Figure 53: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

model for each case.



(a) Original (b) +1 time augmatation (c) Luminosity ×2

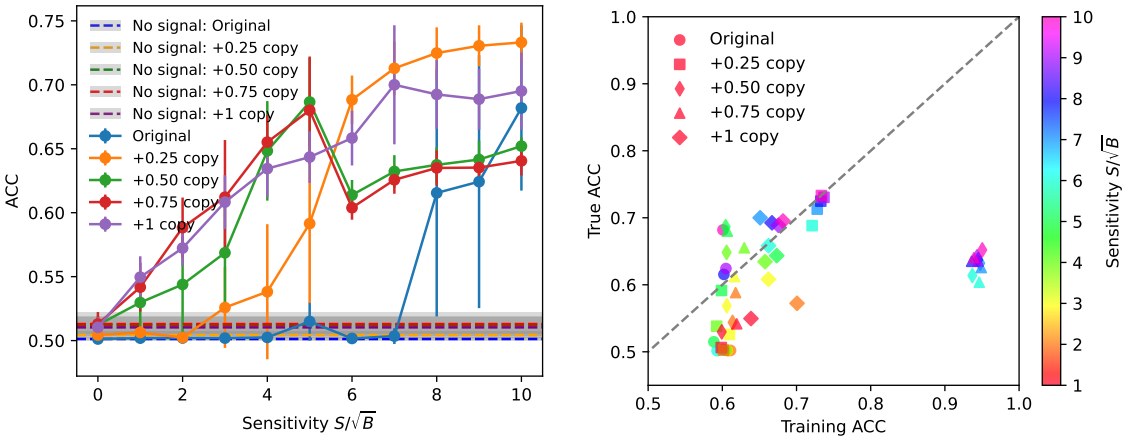
(d) +1 time copy (e) 1 time augmatation (f) half + 1 time copy

Figure 54: The accuracy curves of two different testing sets. We present the accuracy of the best training model. In all cases, they have similar performance.

4.19 Original + x Copy

The training datasets with different ratios of duplicated samples are generated. The training dataset consists of the original sample and x copy sample, where ratio $x = 0.25$, 0.50, 0.75. Figure 55 is the ACC curve and ACC distribution. The training results of dataset adding copy samples are better than the original training sample in the low sensitivity region. The training results differ for $x = 0.50, 0.75$ when $S/\sqrt{B} \geq 6$. From the ACC scatter plot, the model seems to be over-training in this region.

Figure 56 is the sensitivity improvement. For $\varepsilon_{\mathrm{b}} = 10\%$, the $+0.50 \sim +1$ copy samples have similar thresholds lower than the original training sample. Still, the sensitivity improvement is smaller than the original dataset in the high-sensitivity region. When the background efficiency is lower, the +0.50 copy and +0.75 copy samples perform better than other cases, even though they are over-training.



(a) (b)

Figure 55: (a) The performance of CWoLa CNN training with various copy ratio datasets. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

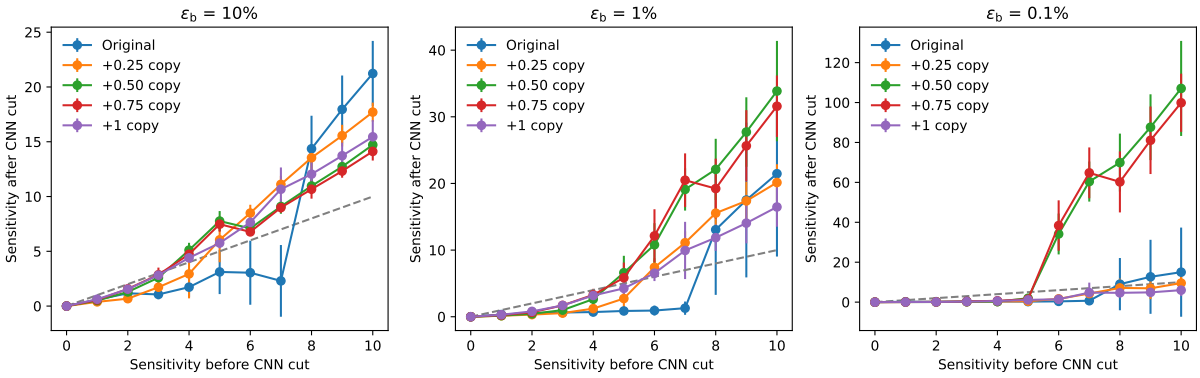
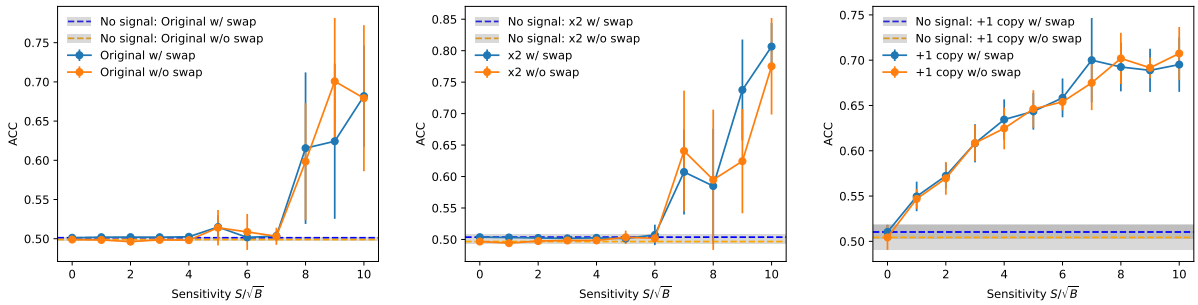


Figure 56: The sensitivities before and after the CWoLa CNN selection. The thresholds are chosen from $\varepsilon_b = 10\%, 1\%, 0.1\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.20 Remove rectification

For a properly constructed classifier, its accuracies should be greater than 50% or it just mislabels the signal and background events. Thus, a couple of codes are used for swapping labels. However, if the dataset causes a larger standard deviation, the swapping procedure might make a fake improvement. To investigate the impact of the "rectification", we remove the code that changes labels and train the CWoLa CNN again. We test on the original, "luminosity ×2" and "+1 copy" datasets. The relabeling process before evaluating the ACC and sensitivity improvement is commented out.

Figure 57 is accuracy curves with different training datasets. Figure 58 is an accuracy scatter plot with various training datasets. The training results with and without swapping procedures are similar for all cases. It seems that the rectification process does not greatly impact the performance. The rectification is not the main reason for the improvement.



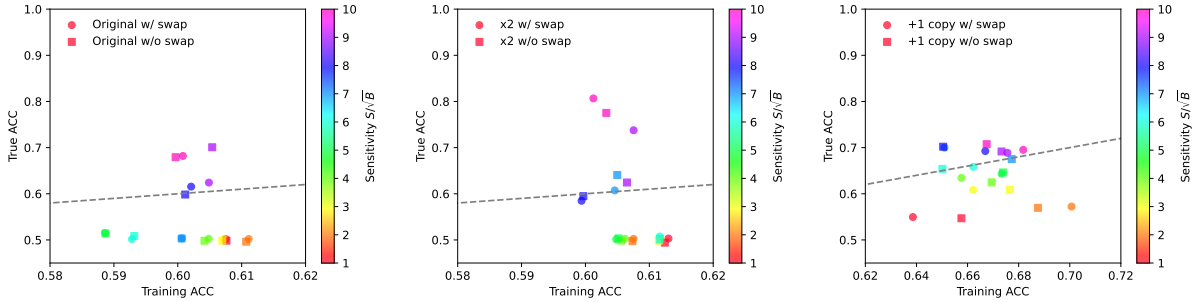
(a) Original (b) Luminosity ×2 (c) +1 time copy

Figure 57: The accuracy curves with and without the rectification procedure. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

Figure 59 is the sensitivity improvement. The training results with and without swapping procedures are similar for all cases. These results are consistent with the results of accuracy curves. The "+1 copy" dataset has the lowest threshold which is between 3 and 4 when $\varepsilon_b = 10\%$, but the sensitivity improvement is smaller than the original and "luminosity ×2" datasets in the high sensitivity region. The "luminosity ×2" threshold is lower than the original sample but higher than the "+1 copy" dataset.

4.21 Compare with Zong-En's results

Figure 60 is Zong-En's results with duplicated samples. The performance of original samples and duplicated samples is consistent.



(a) Original (b) Luminosity $\times 2$ (c) $+1$ time copy

Figure 58: Scatter plots for training ACC and true ACC with and without the rectification procedure. The slope of the grey dashed line is 1, representing the same training and true ACC.

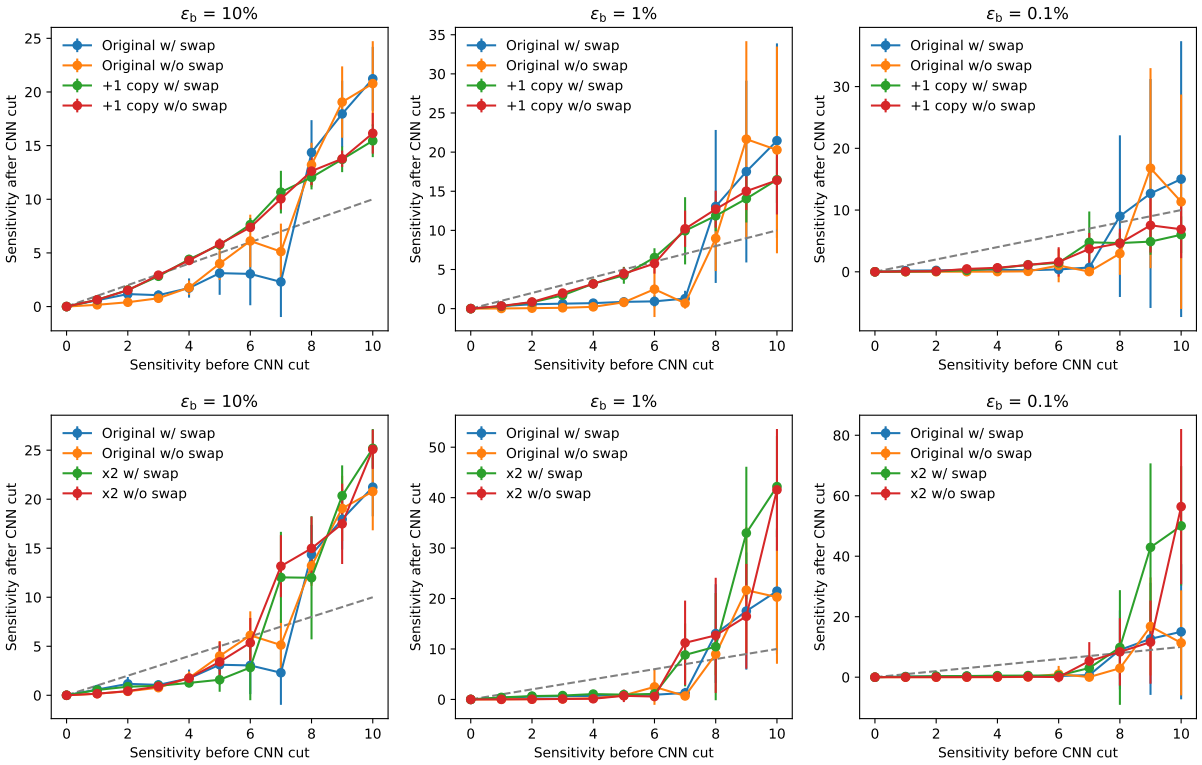


Figure 59: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

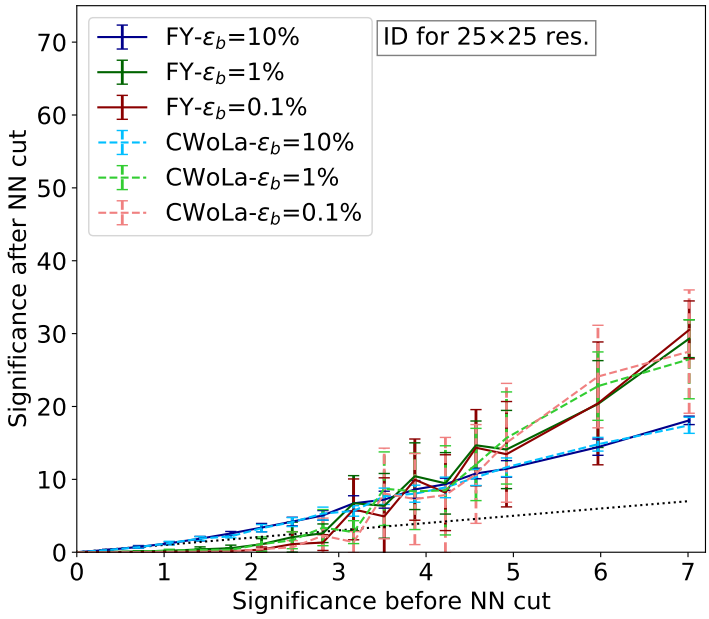


Figure 60: The significance before and after the CWoLa CNN selection. The solid lines are the new results with duplicated samples, and the dashed lines are the previous results. The slope of the dashed grey line is 1, representing the same performance before and after the selection.

Since my training results are inconsistent with Zong-En's, I need to make a comparison with Zong-En's training codes. The main differences are listed below:

1. Model structure: I treated 2 jet images as 2 channels of one diagram. Zong-En passed each jet image through a common CNN and multiplied output values.

2. Training sample preparation: I prepare training samples in the `.npy` file before training. Zong-En uses the generator method to construct samples during the training process.

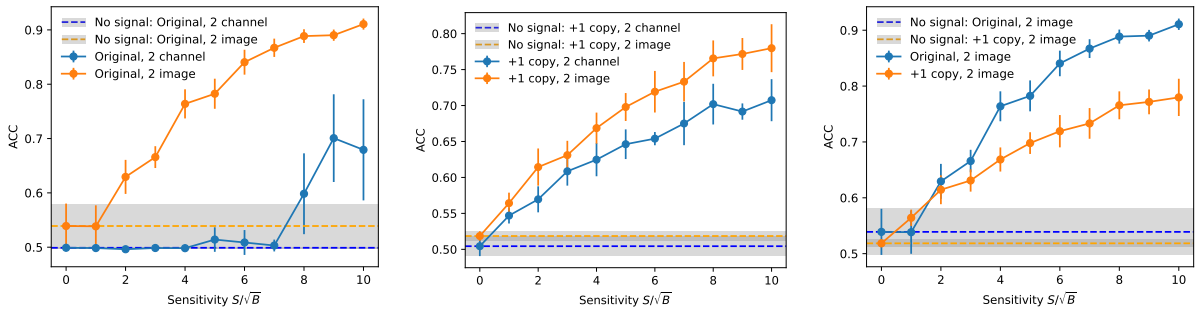
3. Hyperparameter setting: My early stopping patience parameter is 10. Zong-En's patience paramter is set to 30.

4. Tensorflow version: I use Tensorflow 2.11.1. Zong-En uses Tensorflow 2.0.0.

4.21.1 Model structure

Modify the model structure such that the same as Zong-En's model [3]. The main difference is how to process jet images. I treated 2 jet images as 2 channels of one diagram. For Zong-En's model, each jet image is passed through a common CNN and returns a single value. The output of the full network is the product of these two numbers.

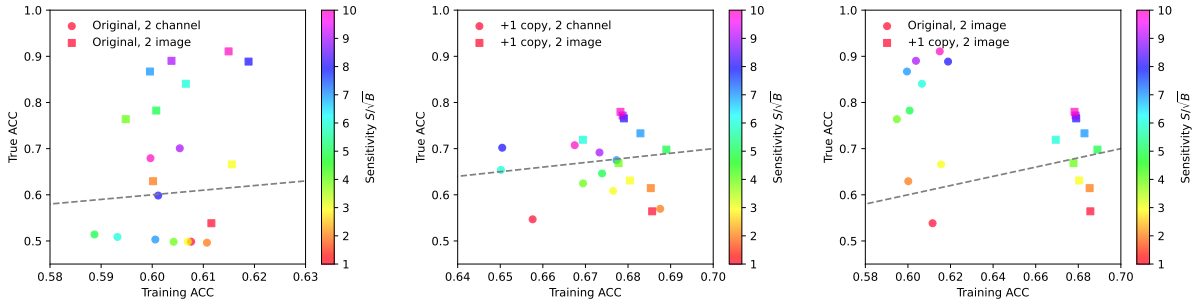
The "2 channel" is my model and the "2 image" is Zong-En's model. Figure 61, 62 are accuracy curves and scatter plots. For the original dataset, the performance of "2 image" is better than "2 channel" in high sensitivity region. Both models perform similarly for the "+1 copy" dataset, but "2 image" still performs slightly better.



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 61: The accuracy curves with different model structures. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

Figure 63 is the sensitivity improvement. The training results of "2 images" are better than "2 channels". The "+1 copy" dataset still has lower thresholds, and the sensitivity



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 62: Scatter plots for training ACC and true ACC with different model structures. The slope of the grey dashed line is 1, representing the same training and true ACC.

improvement is smaller than the original dataset in the high-sensitivity region. The "2 image" model can obtain larger improvement for lower background efficiency, but the threshold is still lower for the "+1 copy" dataset.

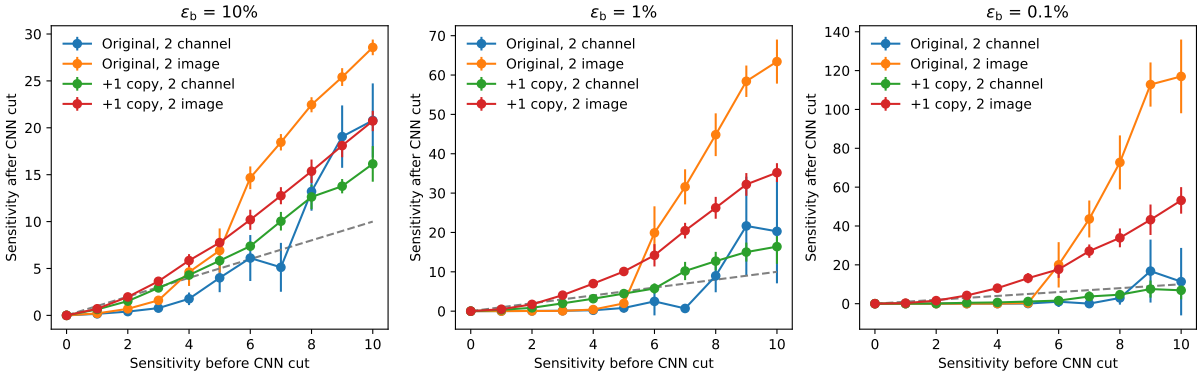
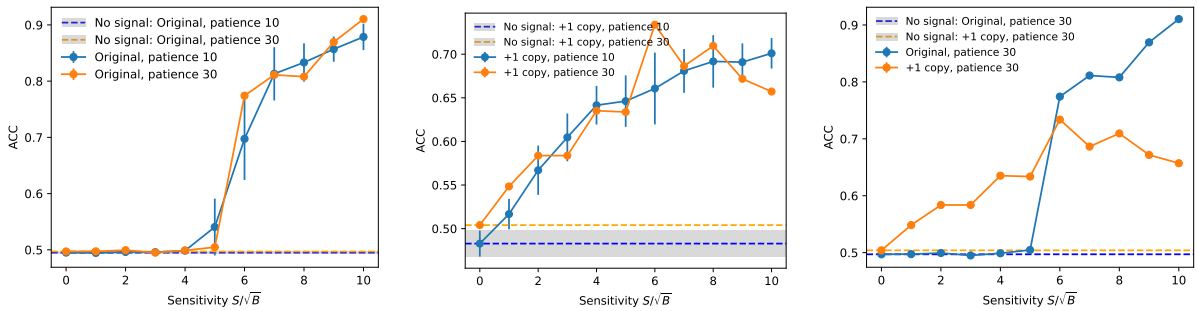


Figure 63: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.21.2 Early stopping patience

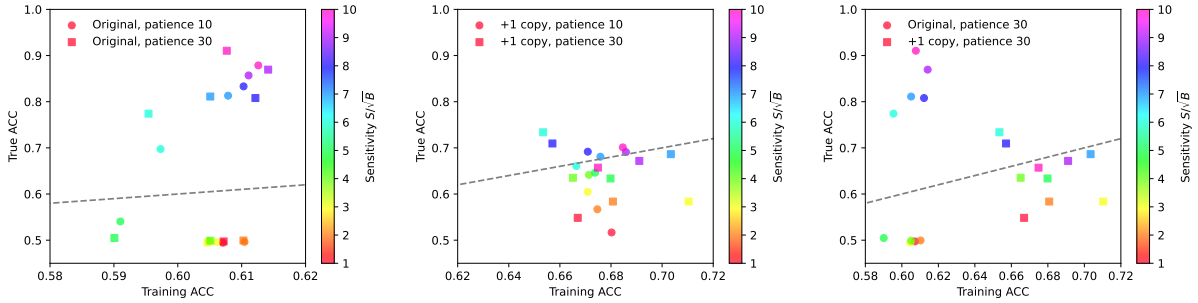
In my previous testing, I always set early stopping patience as 10, but Zong-En set this parameter as 30.

Figure 64, 65 are accuracy curves and scatter plots. The performance "Patience 30" is similar to previous results. It seems that patience 10 is sufficient for our training procedure. Therefore, it does not greatly impact the training performance even though we change patience from 10 to 30.



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 64: The accuracy curves with different early stopping patience. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. Note that we only present the result of one-time training for the "patience 30" case.



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 65: Scatter plots for training ACC and true ACC with different early stopping patience. The slope of the grey dashed line is 1, representing the same training and true ACC.

Figure 66 is the sensitivity improvement. The training results of "patience 10" and "patience 30" are similar. This is consistent with the training accuracy.

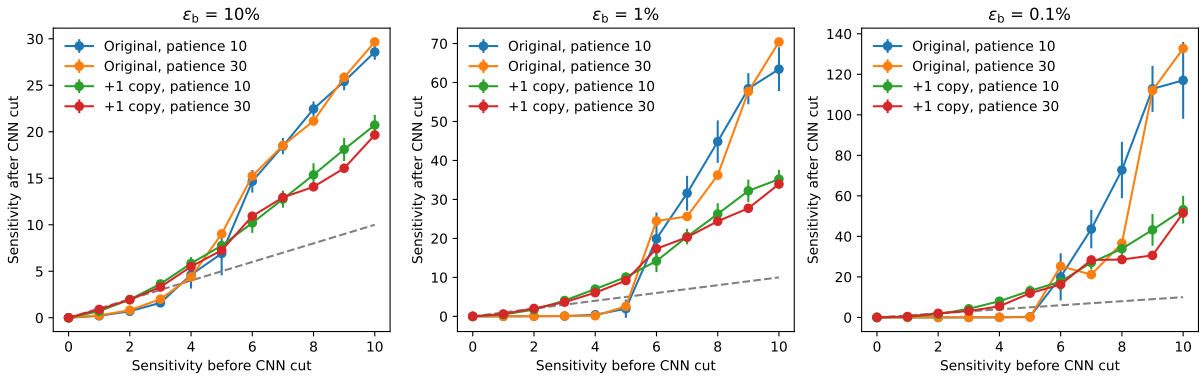


Figure 66: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. Note that we only present the result of one-time training for "patience 30".

4.21.3 Tensorflow version

I use Tensorflow 2.11.1 and Zong-En uses Tensorflow 2.0.0. The difference in version may cause some unexpected issues.

Figure 67, 68 are the training results with Tensorflow 2.5.0. The performance is consistent with version 2.11.1.

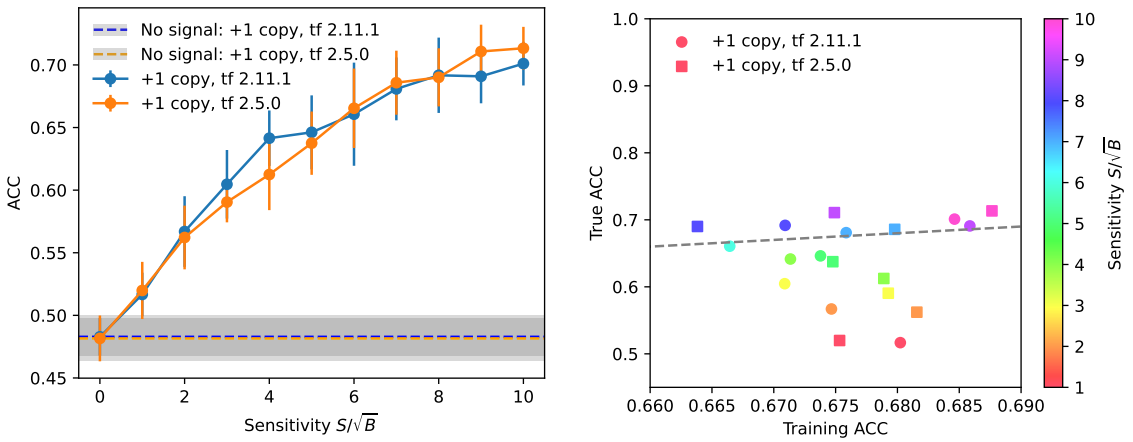
I should have tested my training codes on version 2.0.0, but an error still can not be resolved.

Error messages:

```
...
The tensor cannot be accessed here:
it is defined in another function or code block.
...
```

4.22 Tensorflow Model compile

According to tf.keras.Model.compile documentation:



(a) (b)

Figure 67: (a) The performance of CWoLa CNN training with Tensorflow 2.5.0. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

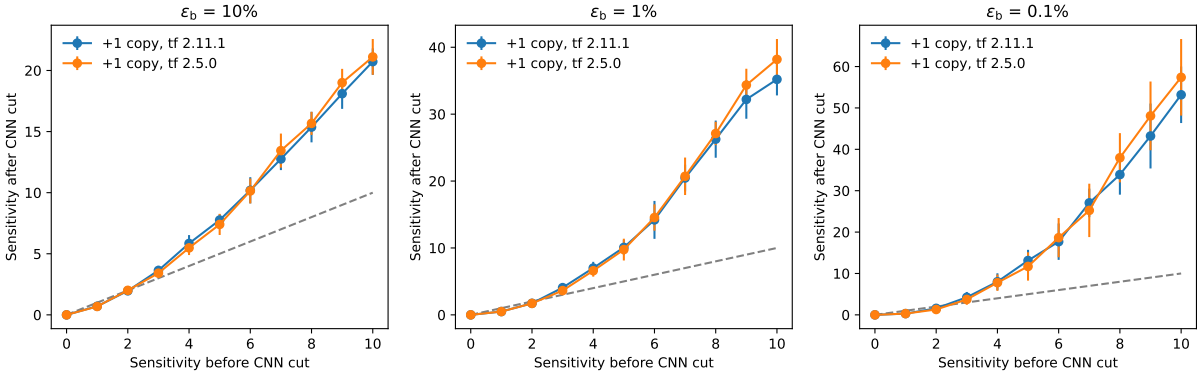


Figure 68: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

> When you pass the strings 'accuracy' or 'acc', we convert this to one of tf.keras.metrics.BinaryAccuracy, tf.keras.metrics.CategoricalAccuracy, tf.keras.metrics.SparseCategoricalAccuracy based on the loss function used and the model output shape.

We modified the model structure and changed the loss function to binary cross entropy. Thus Tensorflow would use the BinaryAccuracy metric, which computes the accuracy with the default threshold of 0.5. However, this accuracy could be better evaluated by threshold scanning.

For the previous model structure, Tensorflow would use CategoricalAccuracy, which checks the output index to see whether it is the same as the index of the true label.

The accuracy results in Figure 64, 65, 67 with the new model structure would be misleading. Only the sensitivity improvement results are meaningful.

4.23 Resolution

To fairly compare with Zong-En's results, the resolution is modified to 25×25. Moreover, we use the significance formula

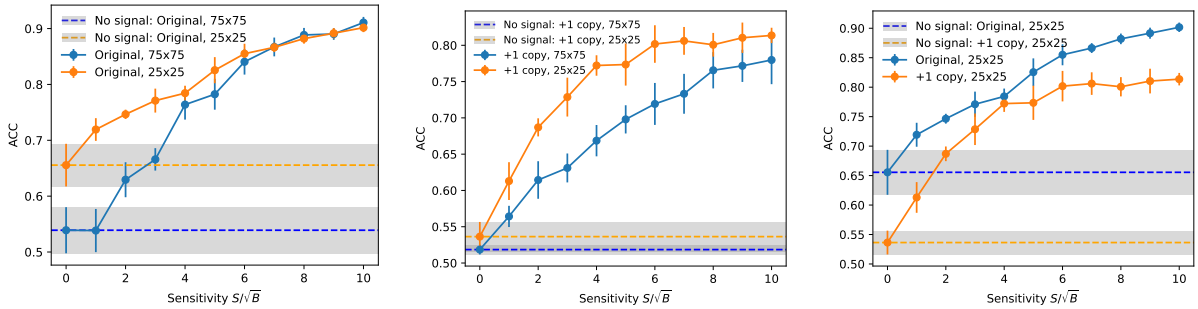
$$\sigma = \sqrt{2\left[(S+B)\log\left(\frac{S}{B}+1\right)-S\right]} \tag{5}$$

where S and B are the numbers of signal and background events. This formula would obtain different results when B is not much greater than S.

Figure 69, 70 are accuracy curves and scatter plots. The performance of "resolution 25×25" is better than "resolution 75×75". The model processing high-resolution images is hard to train because there are more parameters. The training results of the "+1 copy" sample are not better than the original dataset.

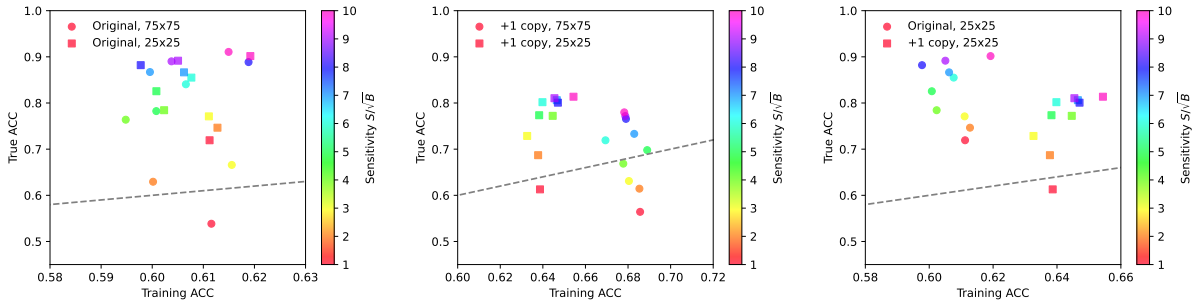
Figure 71 is the sensitivity improvement. The training results of "resolution 25×25" are better than "resolution 75×75". This is consistent with the training accuracy. For $\varepsilon_\mathrm{b}=10\%$, the original dataset with resolution 25×25 has the best performance. For lower background efficiencies, the threshold of the original dataset would be worse than the "+1 copy" dataset by about 1 significance. This difference is smaller than the case of resolution 75×75 (Figure 66).

Figure 72 is my and Zong-En's testing results. All curves seem to have similar results, except the "+1 copy, $\varepsilon_\mathrm{b}=0.1\%$". For Zong-En's results, it can obtain similar performance at high-sensitivity regions. However, in my case, the results are worse in the high-sensitivity areas.



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 69: The accuracy curves with resolution 25×25. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 70: Scatter plots for training ACC and true ACC with resolution 25×25. The slope of the grey dashed line is 1, representing the same training and true ACC.

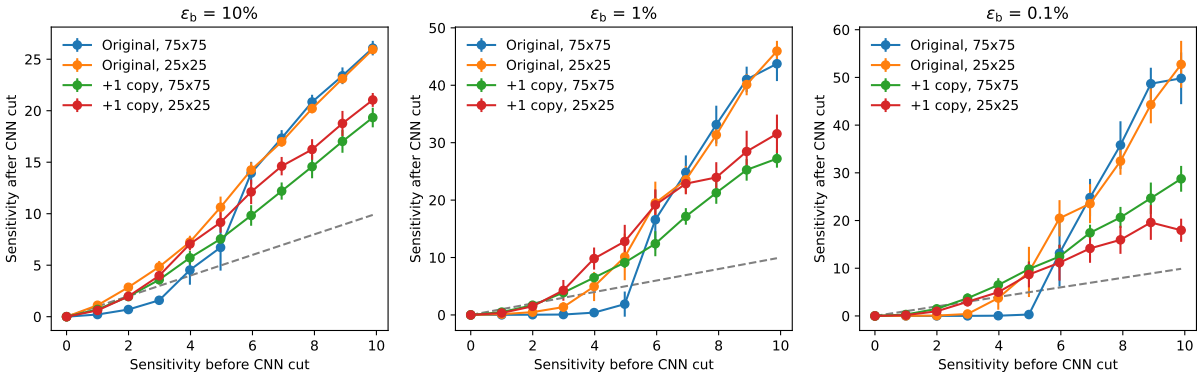
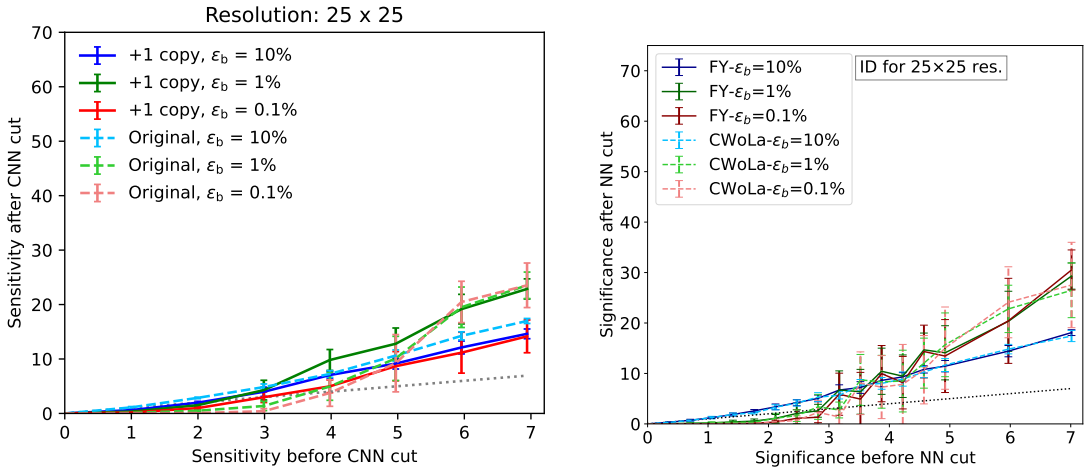


Figure 71: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.



(a) My testing

(b) Zong-En's testing

Figure 72: The significance before and after the CWoLa CNN selection. The solid lines are the new results with duplicated samples, and the dashed lines are the previous results. The slope of the dashed grey line is 1, representing the same performance before and after the selection.

4.24 Other plots

4.24.1 Loss and accuracy across training

Figure 73 and 74 are the loss and accuracy values across training. The training and validation values are evaluated on mixed-label samples, while the testing values are assessed on true-label samples. Thus, the difference between these curves may not be meaningful. The training time of the "+1 copy" dataset is longer than the original samples.

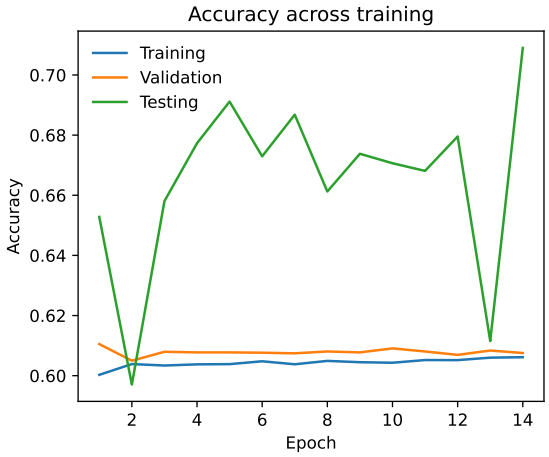
Since the sample size imbalance in signal and sideband regions (Table 15), the accuracy values of the training and validation curve are close to 0.6 but not 0.5.

4.24.2 ROC curve

Figure 75 is ROC curves with different datasets. Here, we plot the ROC curve of the best training model. Resolution 25×25 models perform better for sensitivity $S/\sqrt{B} = 3.0$. The curve of the "+1 copy, 75×75" dataset has weird performance, it does not work as a well-training classifier. The model might overfit the training dataset.

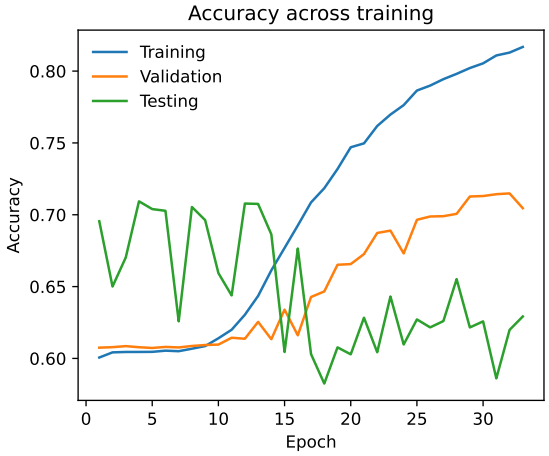
Figure 76 presents ROC curves with various datasets. The "+1 copy" dataset seems to improve the performance at the low-FPR (False Positive Rate) region, while the performance would be worse when we consider the high-FPR region. The model might overfit on some

(a) Original



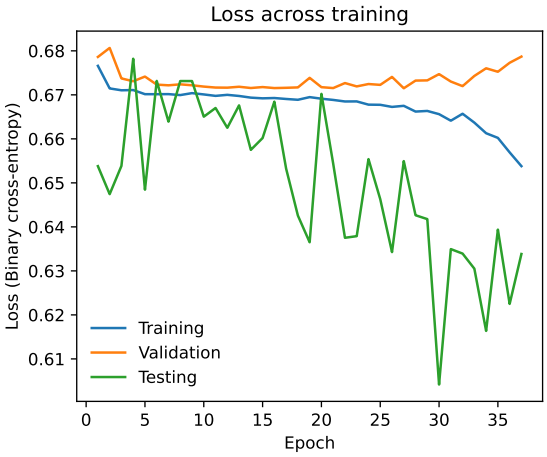
(b) Original

(c) +1 time copy

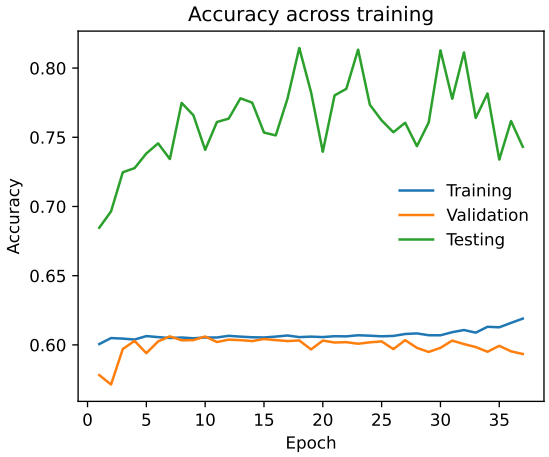


(d) +1 time copy

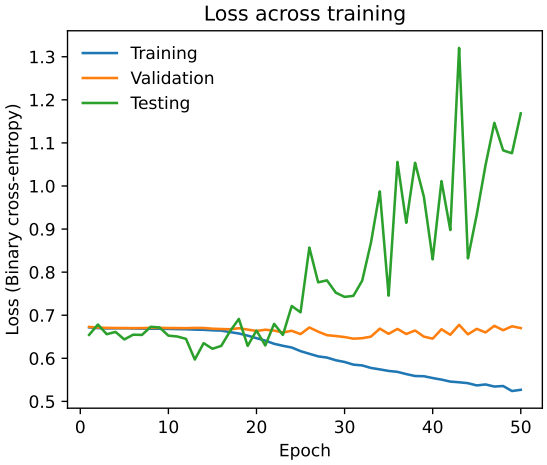
Figure 73: The loss and accuracy values across training. The model is trained on the sensitivity $S/\sqrt{B} = 3.0$ dataset with resolution 75×75. Loss value is evaluated from binary cross-entropy. Accuracy is evaluated with a threshold of 0.5 for training and validation. The testing accuracy is the best accuracy with threshold scanning.



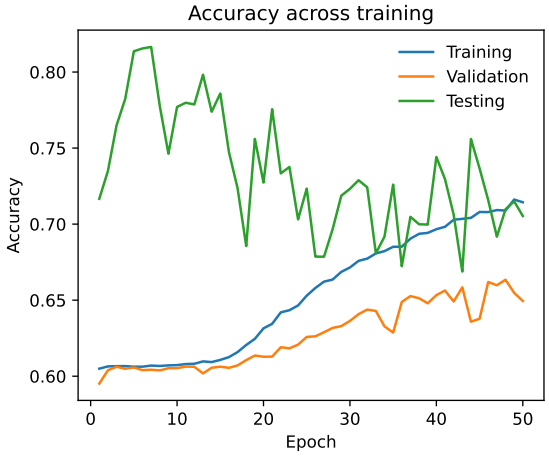
(a) Original



(b) Original

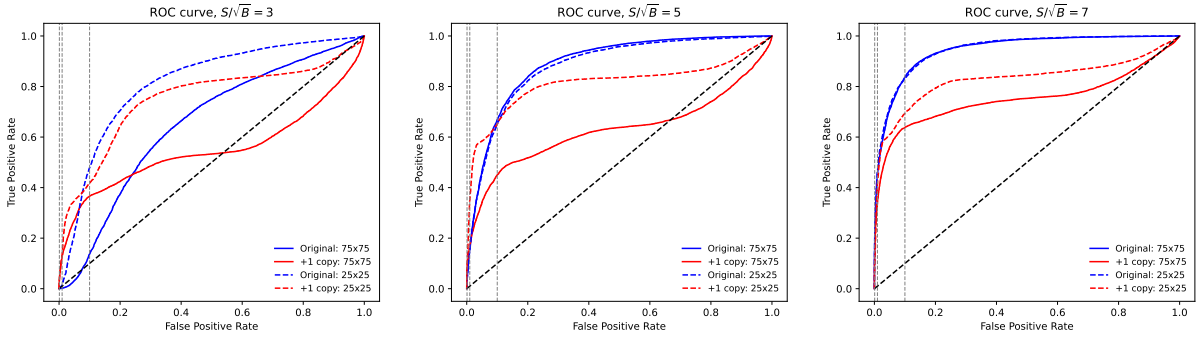


(c) +1 time copy



(d) +1 time copy

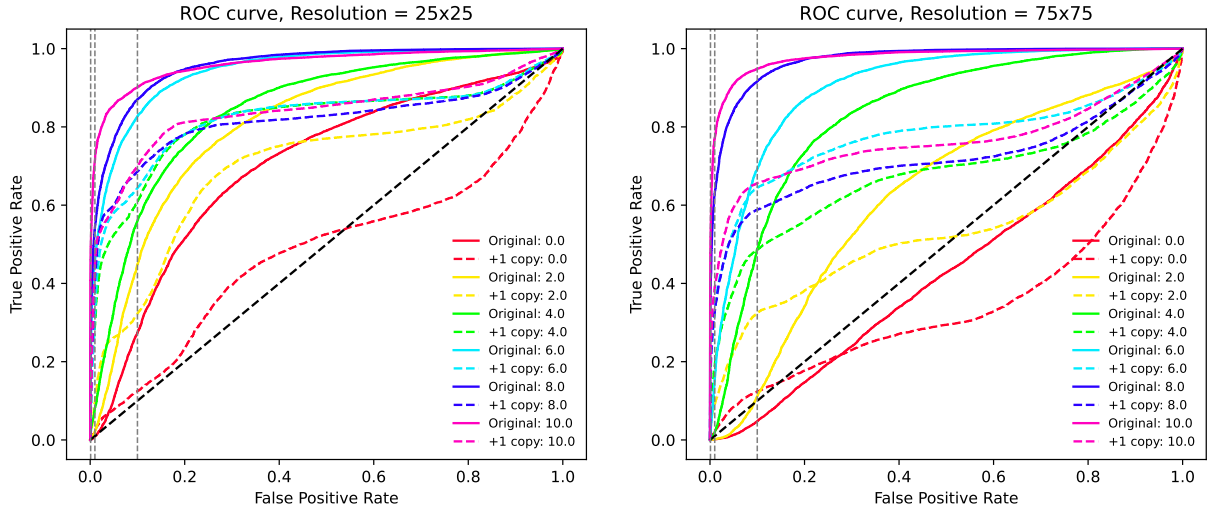
Figure 74: The loss and accuracy values across training. The model is trained on the sensitivity $S/\sqrt{B} = 3.0$ dataset with resolution 25×25. Loss value is evaluated from binary cross-entropy. Accuracy is evaluated with a threshold of 0.5 for training and validation. The testing accuracy is the best accuracy with threshold scanning.



(a) (b) (c)

Figure 75: The grey dashed lines represent $\varepsilon_{\mathrm{b}} = 10\%, 1\%, 0.1\%$. In those plots, the ROC curve is evaluated from the best model.

samples.

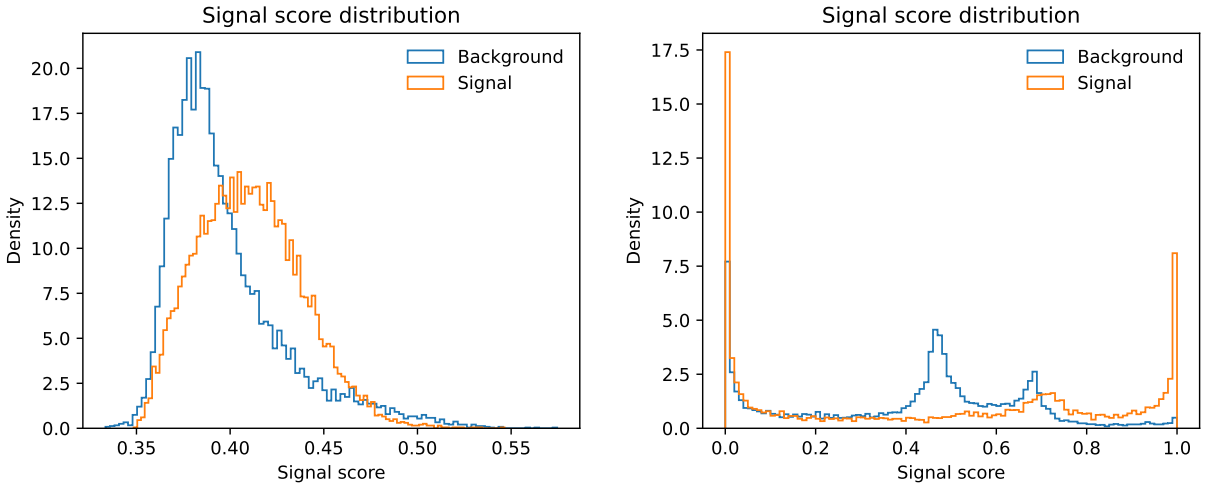


(a) (b)

Figure 76: The grey dashed lines represent $\varepsilon_{\mathrm{b}} = 10\%, 1\%, 0.1\%$. In those plots, the ROC curve is evaluated from the best model.

4.24.3 Signal score

The neural network's output is a number, which is the signal score of an event. Figure 77 is the signal score distribution of the model trained on original datasets and duplicated datasets. These distributions are very different and can explain the ROC curve in Figure 75.



(a) Original

(b) +1 time copy

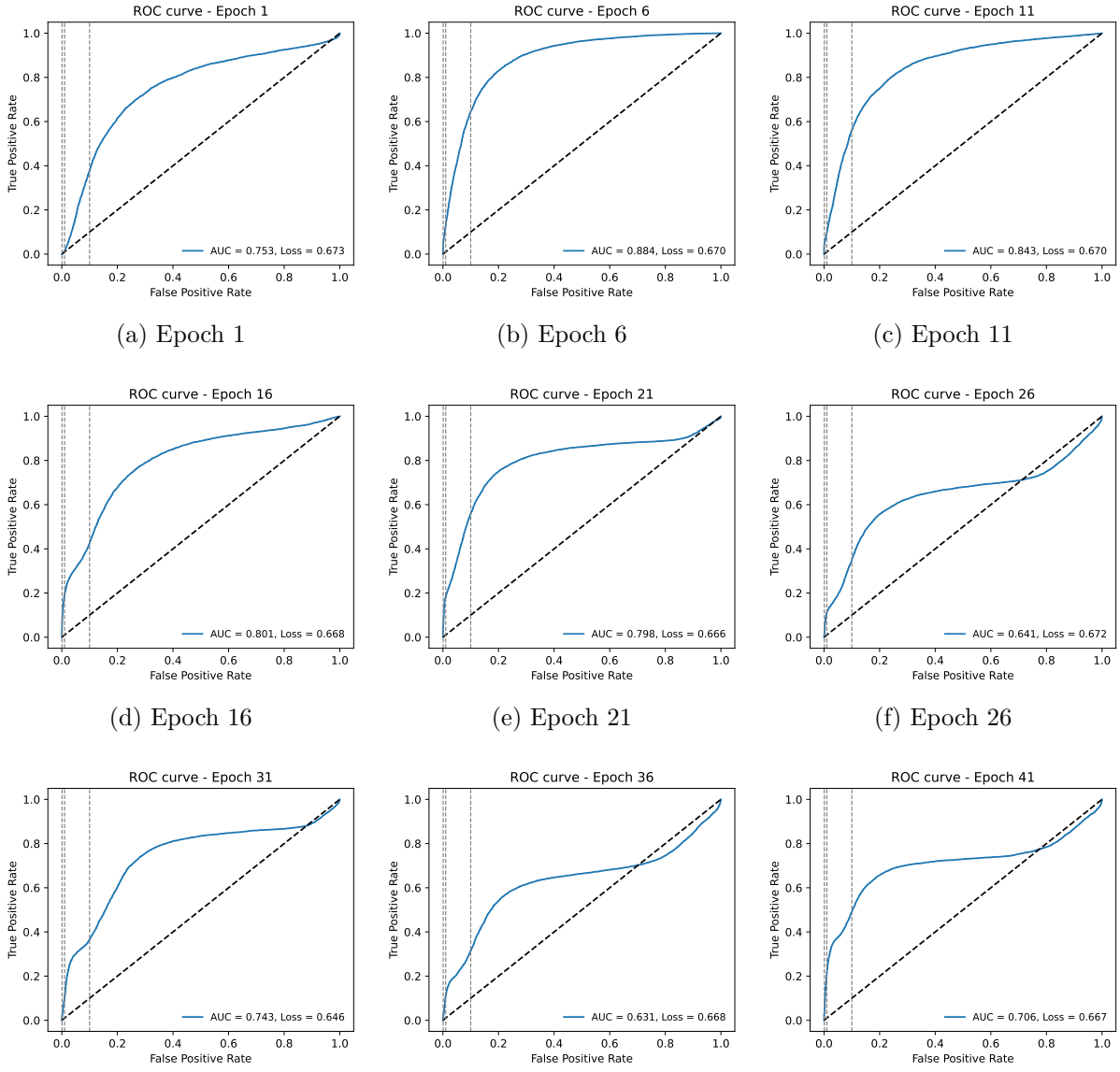
Figure 77: The signal score distribution of the model trained on original and duplicated datasets. The sensitivity $S/\sqrt{B} = 3$. The signal score is evaluated from the best model. The input events are the signal and background events in the signal region.

4.24.4 ROC curve in training process

We found that the training processes are very different for original and duplicated datasets in Figure 74. We plot the ROC curve across the training process to investigate more information in the training process.

Figure 78 is ROC curves at different epochs for duplicated datasets. The ROC curves before Epoch 11 satisfied our normal expectations, but the shape became weird after Epoch 16. Even if the shape or AUC value worsens, the loss value decreases. The training process does not stop at a reasonable stage. It seems that the problem comes from the over-training issue.

Figure 79 is ROC curves at different epochs for original datasets. The ROC curves before Epoch 21 look normal. The shape of the ROC curve is a little deformed after Epoch 26, but the impact is minimal compared with the "+1 copy" case. After the deformation of the ROC curve is observed, the training process does not take long to stop, which is terminated at the reasonable stage. It seems that there is no over-training issue for the original training sample.

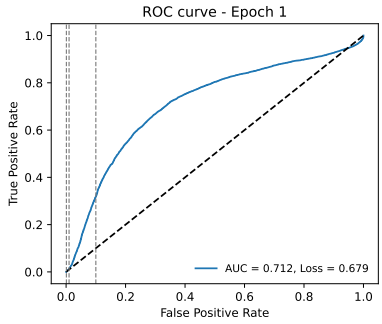


(a) Epoch 1 (b) Epoch 6 (c) Epoch 11

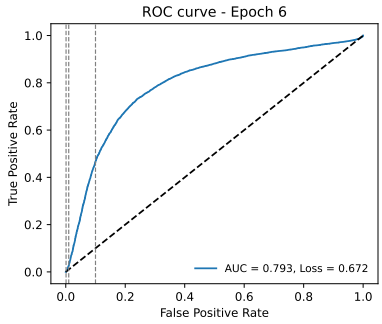
(d) Epoch 16 (e) Epoch 21 (f) Epoch 26

(g) Epoch 31 (h) Epoch 36 (i) Epoch 41

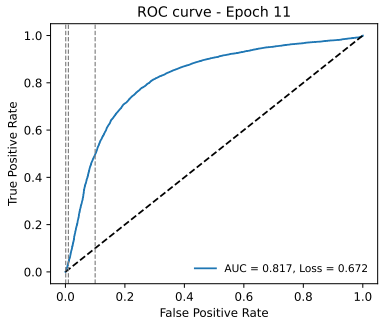
Figure 78: These ROC curves are plotted from true label samples in the signal region. The model is trained on the duplicated dataset with sensitivity $S/\sqrt{B} = 3$. The loss value is the validation loss in the training process. The grey dashed lines represent $\varepsilon_{\mathrm{b}} = 10\%, 1\%, 0.1\%$.



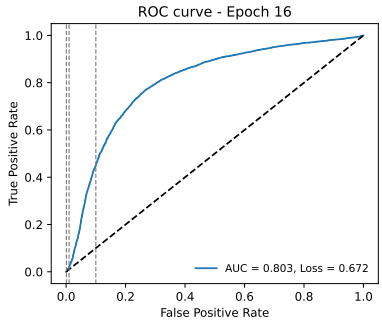
(a) Epoch 1



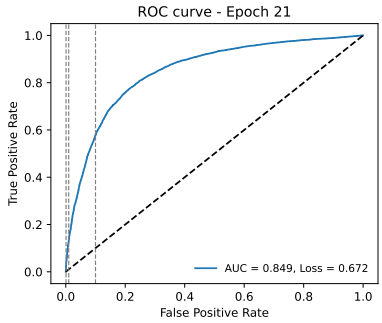
(b) Epoch 6



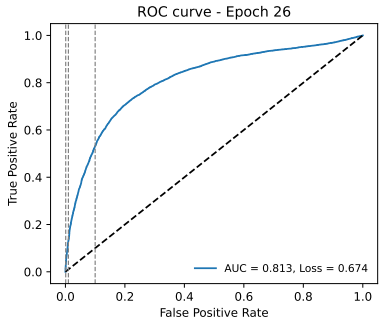
(c) Epoch 11



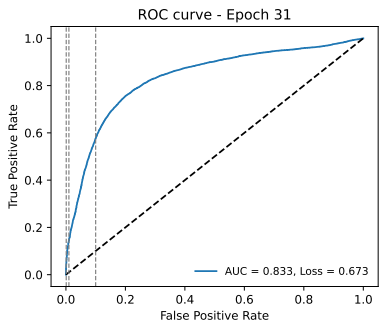
(d) Epoch 16



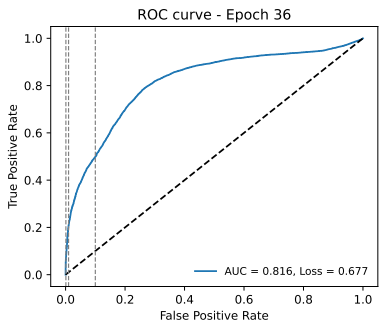
(e) Epoch 21



(f) Epoch 26



(g) Epoch 31

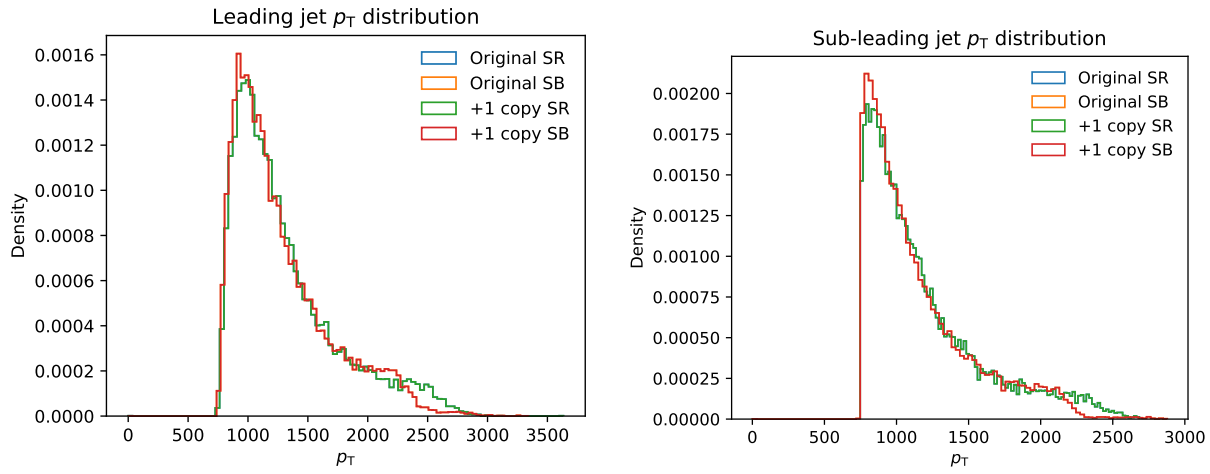


(h) Epoch 36

Figure 79: These ROC curves are plotted from true label samples in the signal region. The model is trained on the original dataset with sensitivity $S/\sqrt{B} = 3$. The loss value is the validation loss in the training process. The grey dashed lines represent $\varepsilon_{\mathrm{b}} = 10\%, 1\%, 0.1\%$.

4.25 Label in .npy file

To make sure the labeling in the duplicated dataset is correct, we select the signal and background events from the original and duplicated datasets and plot some distributions. Figure 80 is p_{T} distributions. Figure 81 is average images. They all have the same results. It seems that the labeling of the duplicated dataset is correct.



(a) Leading jet

(b) Sub-leading jet

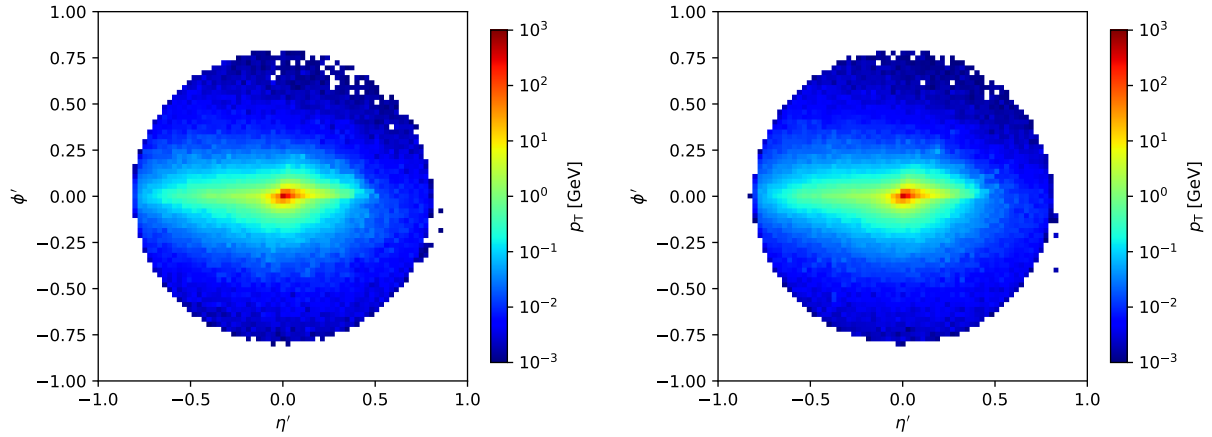
Figure 80: The p_{T} distributions for original and duplicated datasets. The p_{T} is the sum of the transverse momentum in each pixel. The differences of each bin in the corresponding region are all equal to zero. Therefore, they have the same distribution.

4.26 More testing on new model structure

Because we have modified the neural network model structure and the resolution of jet images, we would like to test the "+1 augmentation" and "luminosity ×2" datasets with these new settings. Then, we can make similar plots as Figure 37 and 38.

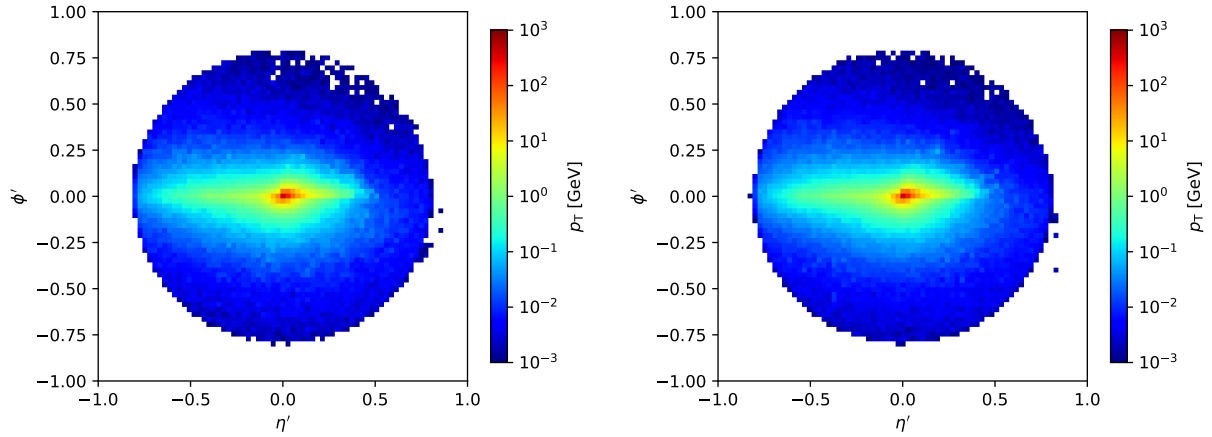
Figure 82 is the ACC curve and ACC distribution. The original and "luminosity ×2" datasets perform similarly. The "+1 augmentation" and "+1 copy" datasets perform similarly and perform worse than the original and "luminosity ×2" samples. Since the training ACC of the "+1 augmentation" and "+1 copy" datasets is slightly higher than the original dataset, they may encounter the over-training problem.

Figure 83 is the sensitivity improvement. The performance of the "luminosity ×2" dataset is similar to the original datasets. "Luminosity ×2" can improve sensitivity using lower background efficiency. The "+1 augmentation" performs similarly to the "+1 copy" case. It exhibits better results in high-sensitivity regions compared with "+1 copy" datasets.



(a) Original, SR

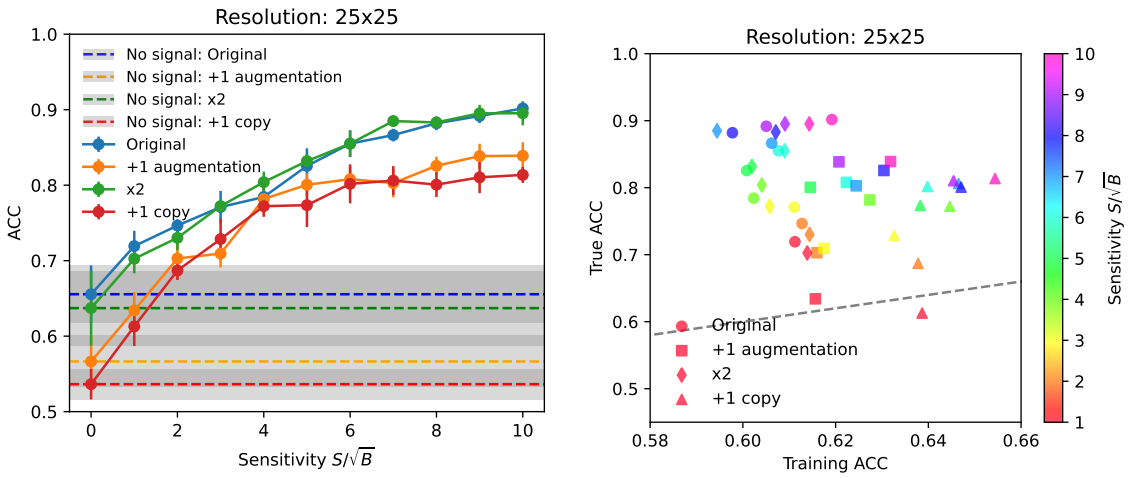
(b) Original, SB



(c) Duplicated, SR

(d) Duplicated, SB

Figure 81: The average jet images in signal and sideband region. The jet images are similar for original and duplicated datasets. The difference between two jet images in the corresponding region is exactly zero. Therefore, they have the same distribution.



(a) (b)

Figure 82: (a) The performance of CWoLa CNN training with different samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

These results are similar to Figure 71. For $\varepsilon_b = 10\%$, the original and "luminosity $\times 2$" datasets have the best performance. For lower background efficiencies, their thresholds would be worse than the "+1 copy" and "+1 augmentation datasets.

4.27 Modify the sideband region

We found the training accuracy is close to 0.6 but not 0.5 in Figure 74. The reason is the imbalance of the data size. The size of sideband region samples is larger than that of signal region samples. To investigate the impact of the sideband region, the sideband boundary is modified from $[4300, 4700] \cup [5500, 5900]$ GeV to $[4400, 4700] \cup [5500, 5800]$ GeV.

After we modified the sideband region, the training sample sizes of each category are summarized in Table 17.

Figure 84 is the loss and accuracy values across training with the new sideband region. The training and validation values are evaluated on mixed-label samples, while the testing values are assessed on true-label samples. The training accuracy is close to 0.52, which satisfied our expectations. The "+1 copy" dataset takes longer to finish the training process. The "+1 copy" dataset seems to have an over-training issue.

Figure 85, 86 are accuracy curves and scatter plots. The sideband region does not highly impact the training performance.

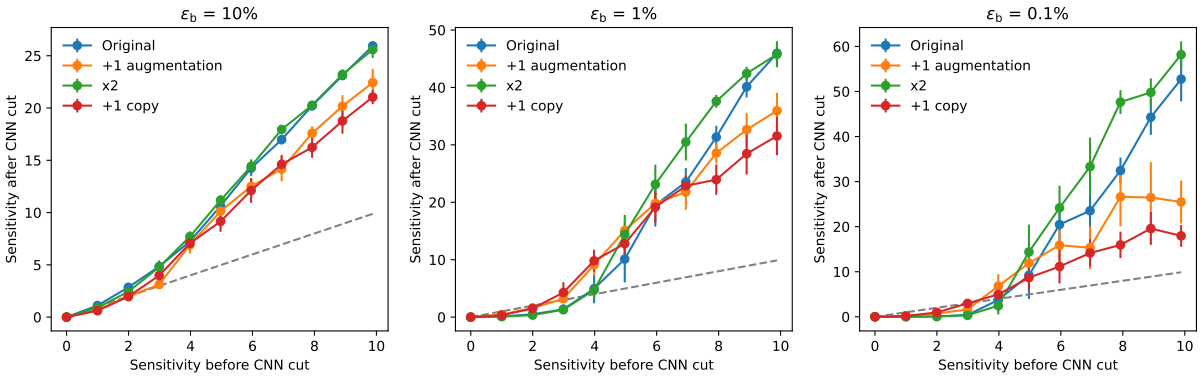
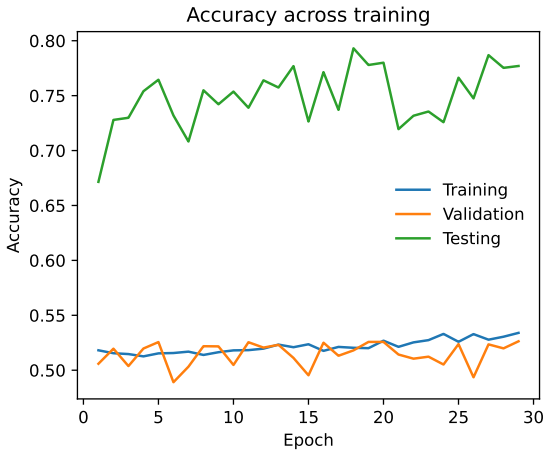


Figure 83: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

Table 17: The training sample size for the mixed sample. We set sensitivity $S/\sqrt{B} = 1$ in the signal region and evaluate the number of events in the signal region. The sideband region is modified to $[4400, 4700] \cup [5500, 5800]$ GeV.

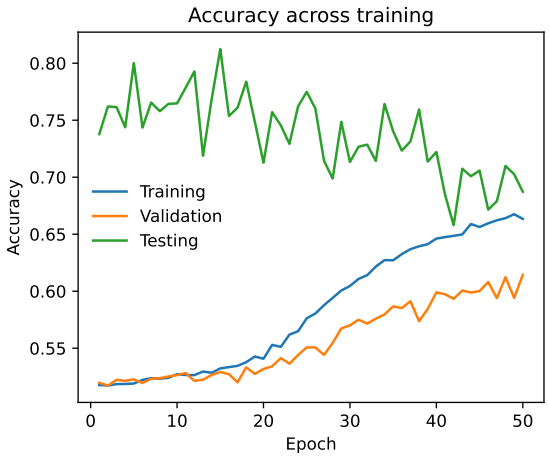
	True label	
Mixed sample	Signal	Background
Signal region	138	19k
Sideband region	34	20k

(a) Original



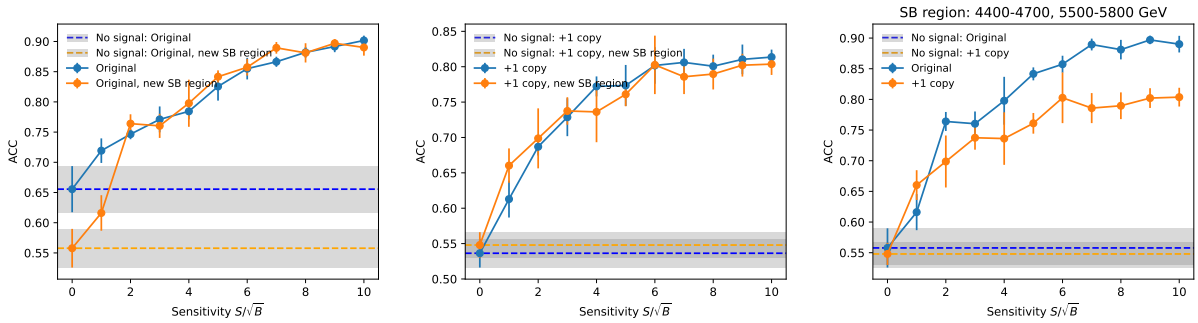
(b) Original

(c) +1 time copy



(d) +1 time copy

Figure 84: The loss and accuracy values across training. The model is trained on the sensitivity $S/\sqrt{B} = 3.0$ dataset with resolution 25×25. The sideband region is modified to $[4400, 4700] \cup [5500, 5800]$ GeV. Loss value is evaluated from binary cross-entropy. Accuracy is evaluated with a threshold of 0.5 for training and validation. The testing accuracy is the best accuracy with threshold scanning.



(a) Original (b) +1 time copy (c) Original v.s. Copy

Figure 85: The accuracy curves with resolution 25×25 and the new sideband region. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

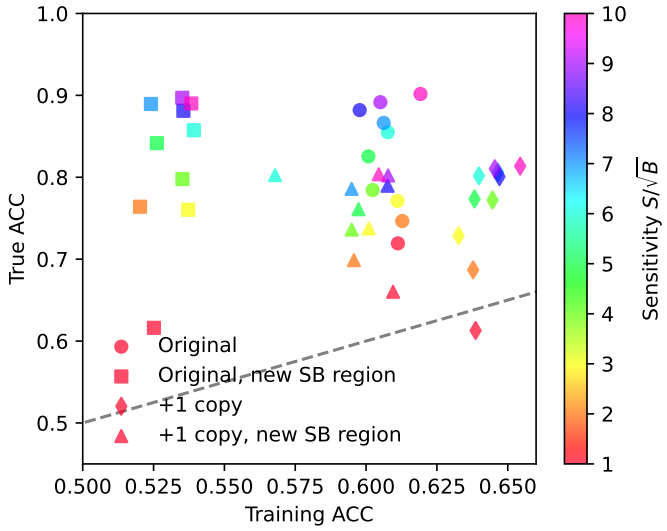


Figure 86: Scatter plots for training ACC and true ACC with resolution 25×25 and the new sideband region. The slope of the grey dashed line is 1, representing the same training and true ACC.

Figure 87 is the sensitivity improvement. The results are consistent with the training accuracy. Even though we modified the sideband region, the training performance is similar to the previous one. The original dataset perform best for $\varepsilon_b = 10\%$. For lower background efficiencies, the threshold of the original dataset would be worse than the "+1 copy" dataset about 1 significance.

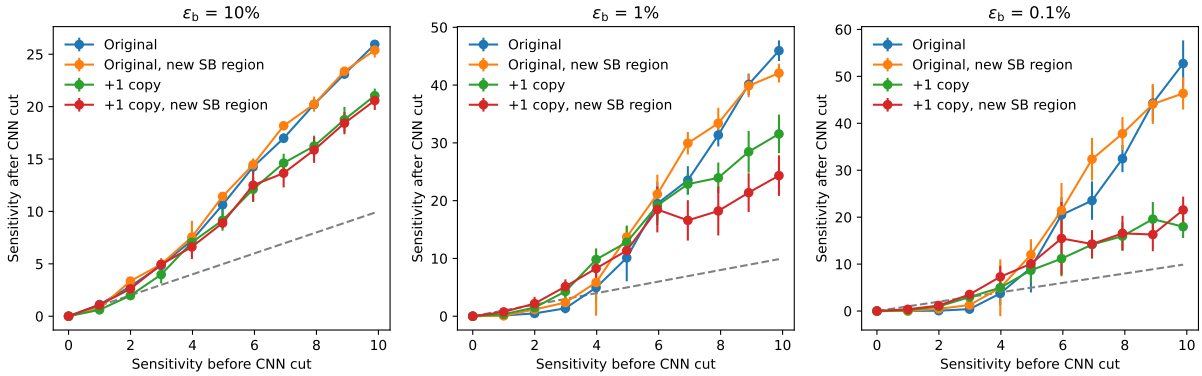


Figure 87: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.28 Modify the procedure of preparing validation dataset

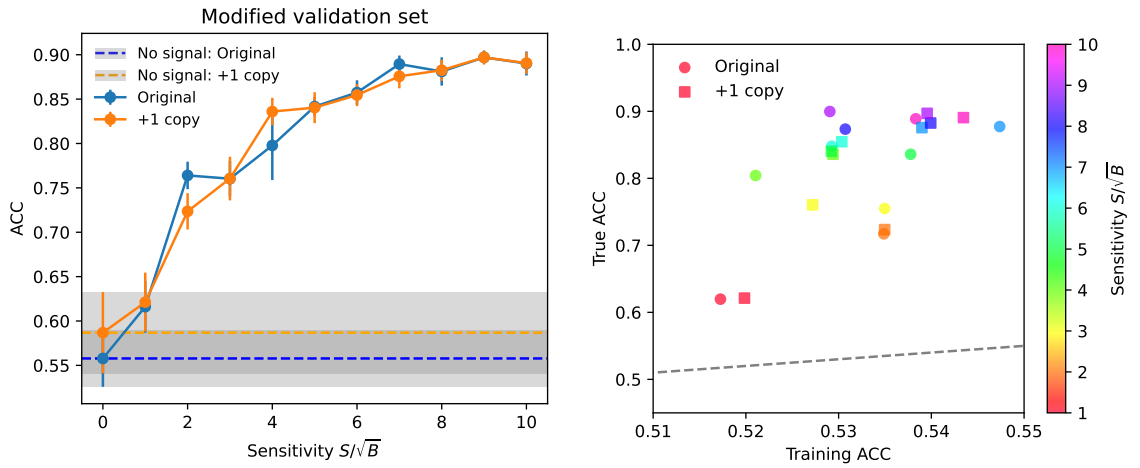
We found the training process can not stop at a reasonable stage in Figure 78. Since the early stopping technique utilizes the validation loss to determine when the training process should be stopped, the over-training issue may come from the validation dataset.

The validation dataset is prepared at the training stage. The `.npy` file would be split into the training and validation datasets with ratios of 0.8 and 0.2, respectively. This procedure works fine for original datasets. However, it is possible that the training and validation datasets could have some common samples for duplicated datasets. The splitting process randomly chooses samples from `.npy` file.

We change the procedure for preparing validation datasets to prevent some samples in the training and validation datasets. First, we split the validation set from the original dataset. Then, we duplicate samples to make the "+1 copy" datasets. In this way, the training and validation dataset would not contain common samples.

Figure 88 is accuracy curves and the scatter plot. The training results are similar for original and duplicated datasets.

Figure 89 is the sensitivity improvement. The results are consistent with the training accuracy. The original and duplicated datasets perform similarly.



(a) (b)

Figure 88: (a) The performance of CWoLa CNN training with different samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

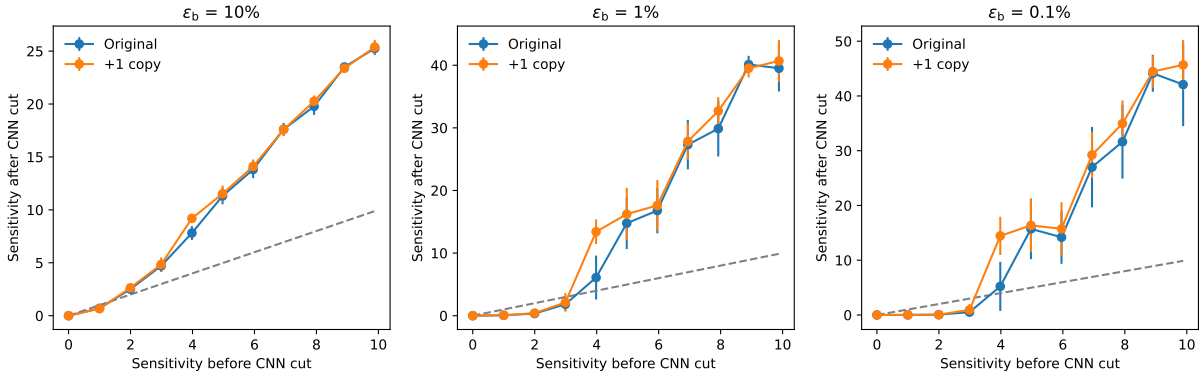


Figure 89: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.29 New data process flow

1. Generate the sample file in `.root` format. Following Section 4.1.
2. Apply the selection cuts described in Section 4.3 and save the event passing the cuts in `HDF5` format. Note that the sideband region is modified to $[4400, 4700] \cup [5500, 5800]$ GeV. The file contains the information listed below
   - The $(p_{\mathrm{T}}, \eta, \phi)$ of leading and sub-leading jet constituents.
   - Total invariant mass m_{jj}.
   - Type of event: 1 for signal, 0 for background.
3. Make mixed sample in `HDF5` format. Following Section 4.5, we can compute the size of datasets. 80% for training set, 20% for validation set.
4. (Optional) Apply data augmentation in `HDF5` format training dataset. Following Section 4.9.
5. Generate the jet image from `HDF5` data and save in `.npy` file.

The key difference with Section 4.8 is step 3. To prevent a similar (common) sample in the training and validation set, the dataset would split into training and validation sets first, then the data augmentation is only applied to the training set.

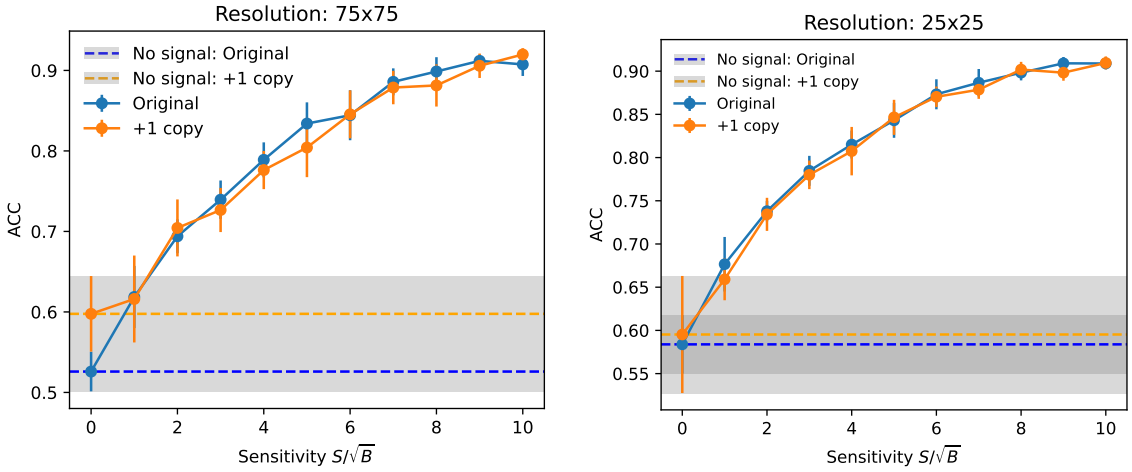
Figure 90 is accuracy curves. We verify that the training results are similar for original and duplicated datasets. Figure 91 is accuracy curves with augmented samples. The training results are also similar for original and augmented datasets. The augmented samples do not improve the training results.

Figure 92 is the sensitivity improvement. The results are consistent with accuracy curves. The original, duplicated, and augmented datasets all perform similarly.

4.30 Enlarge the data size with more simulated samples

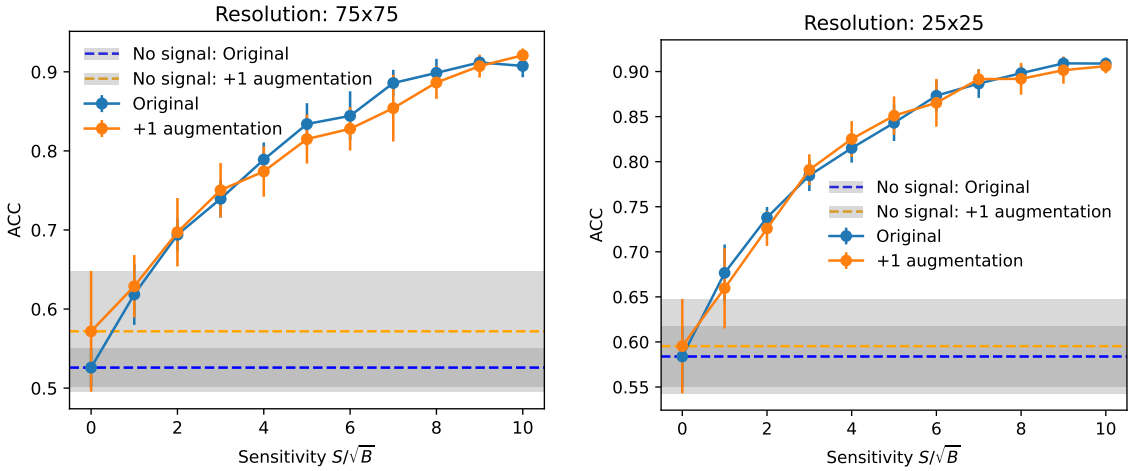
We enlarge the training data size with more simulated samples to identify the upper limit of augmented datasets. More specificaly, the luminosity is scaled to $\mathcal{L} = 139 \times 2$ fb^{-1}.

Figure 93 shows the sensitivity improvement. The "luminosity $\times 2$" datasets perform slightly better than the original dataset. As the background efficiency decreases, the improvement becomes more pronounced. This suggests there is room for improvement in training performance with augmented samples. Note that if we want to compare the model trained on a similar signal sample size, we should compare the "+1 augmentation" results with the



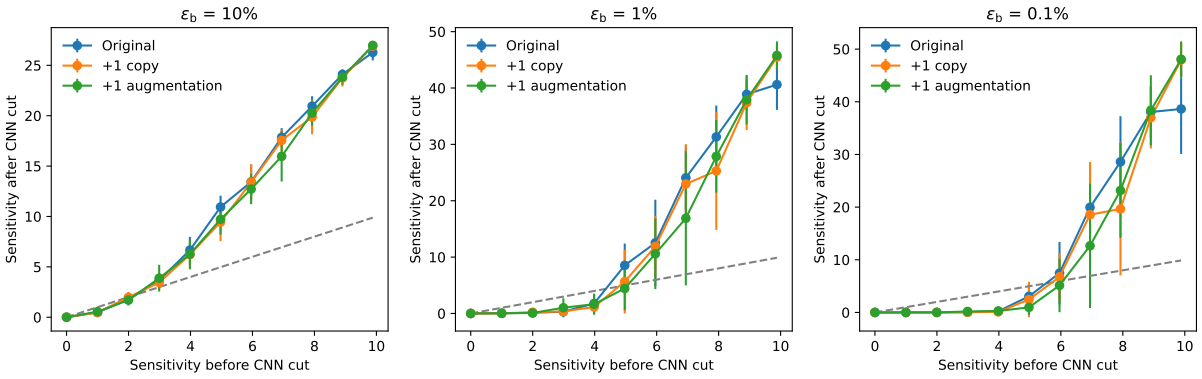
(a) (b)

Figure 90: The performance of CWoLa CNN training with different samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

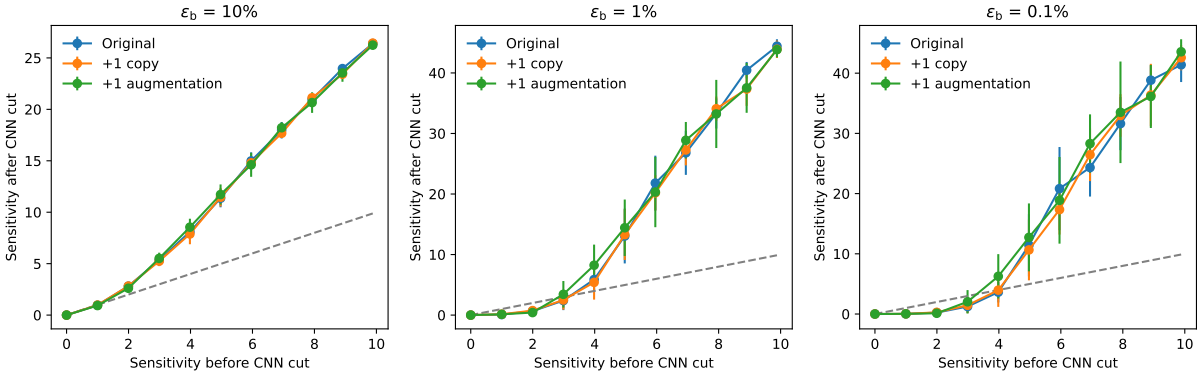


(a) (b)

Figure 91: The performance of CWoLa CNN training with different samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.



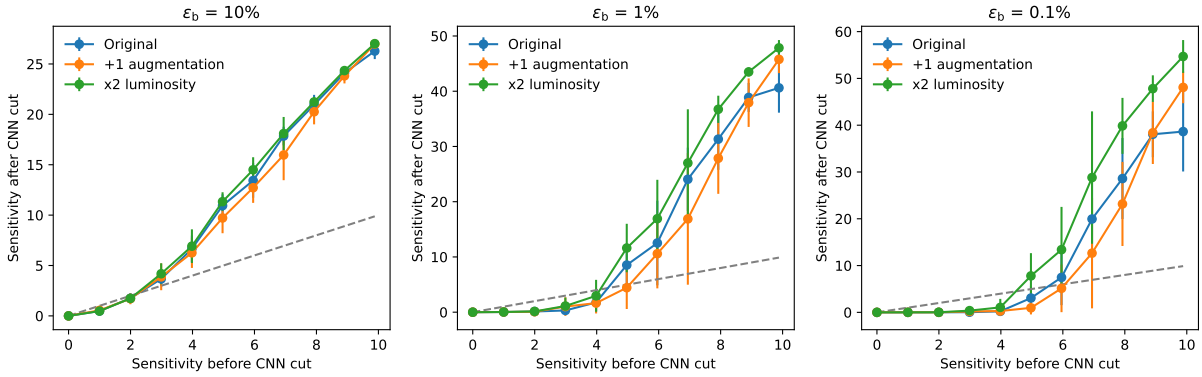
(a) Resolution: 75×75



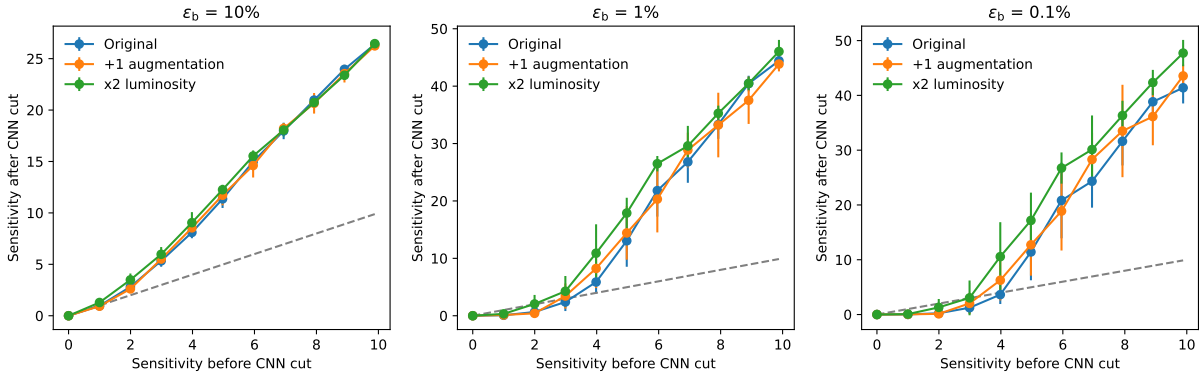
(b) Resolution: 25×25

Figure 92: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

point on the "luminosity $\times 2$" curve corresponding to $\sqrt{2}$ times the sensitivity, similar to Section 4.10.



(a) Resolution: 75×75



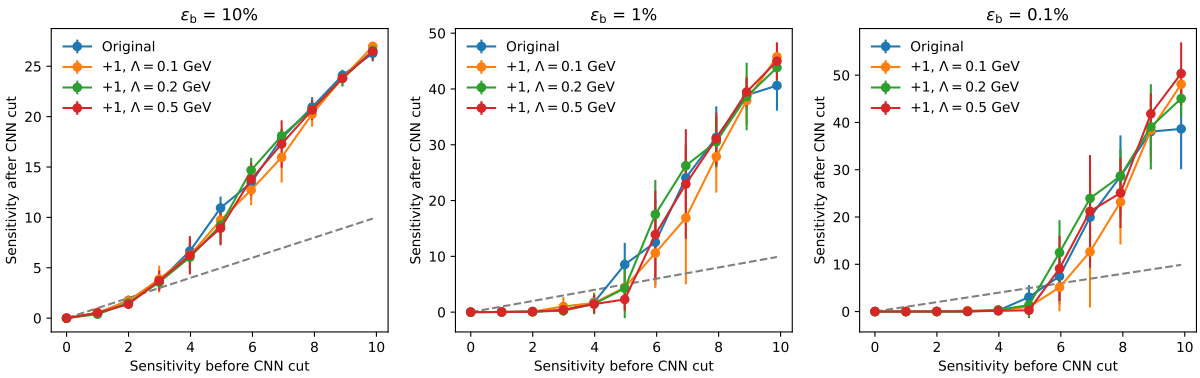
(b) Resolution: 25×25

Figure 93: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

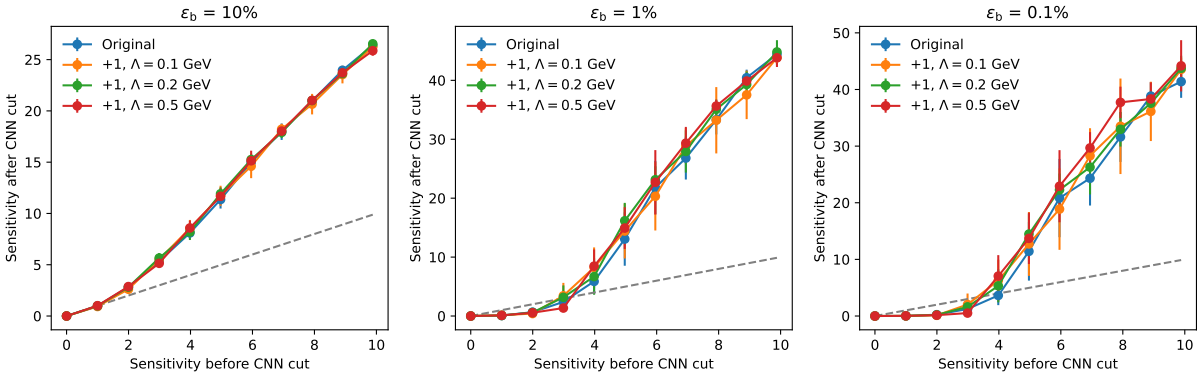
4.31 Smearing scale in $\eta - \phi$ smearing

To improve the performance of $\eta - \phi$ augmentation, we generate samples with various Λ for CWoLa CNN training. We test $\Lambda = 200, 500$ MeV samples.

Figure 94 shows the sensitivity improvement with various smearing scales. The training results are similar for all cases. The $\eta - \phi$ augmentation seems not to improve the training performance.



(a) Resolution: 75×75



(b) Resolution: 25×25

Figure 94: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. Here, the "+1" means "+1 augmentation" datasets.

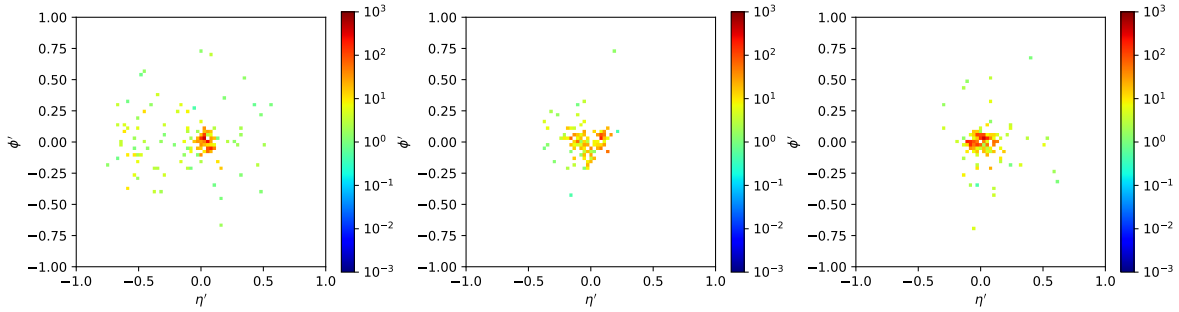
4.32 Implement p_{T} smearing

Similar to the Section 3.2, we apply the p_{T} smearing on our training sample. Specifically, the transverse momentum p_{T} of each jet constituent is resampled according to a Normal distribution centered on the original value with a standard deviation $f(p_{\mathrm{T}})$

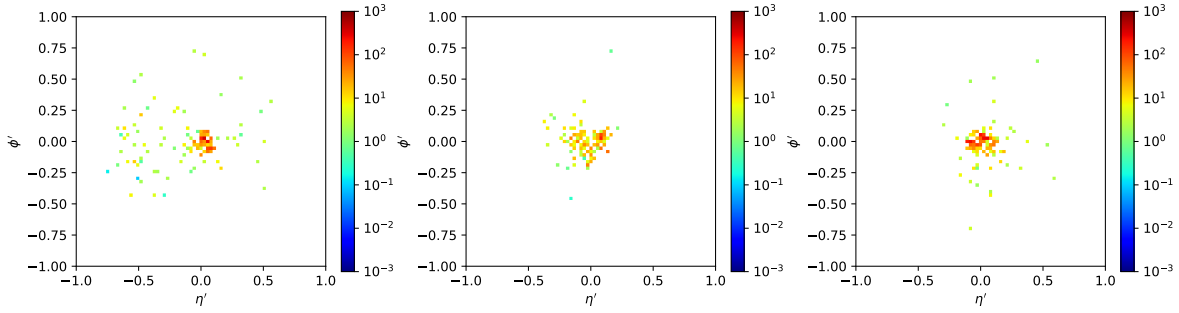
$$p'_{\mathrm{T}} \sim \mathcal{N}\left(p_{\mathrm{T}}, f(p_{\mathrm{T}})\right), \quad f(p_{\mathrm{T}}) = \sqrt{0.052 p_{\mathrm{T}}^2 + 1.502 p_{\mathrm{T}}} \tag{6}$$

where p'_{T} is the augmented transverse momentum, $f\left(p_{\mathrm{T}}\right)$ is the energy smearing applied by `Delphes` (the p_{T}'s are normalised by 1 GeV). Note that if a constituent has negative p'_{T}, this jet constituent would be dropped.

Figure 95 is the jet image before and after the p_{T} augmentation. These jet images look similar but not the same.



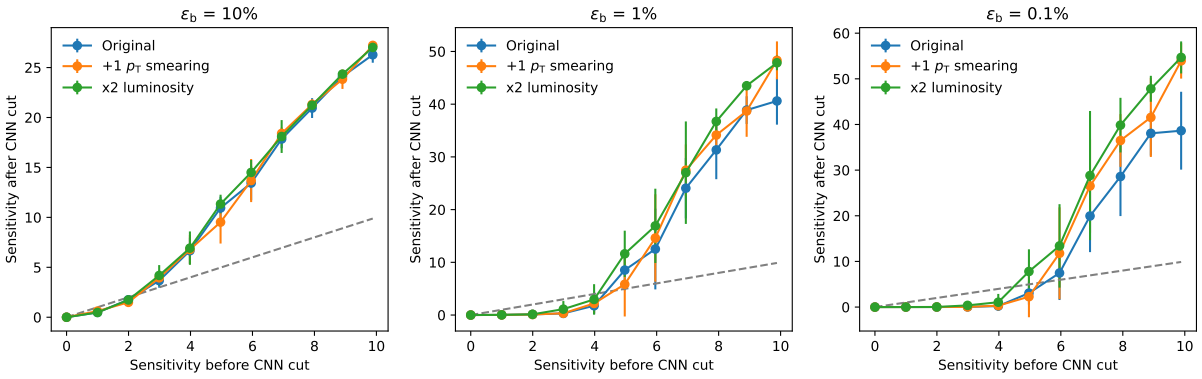
(a) Before augmentation



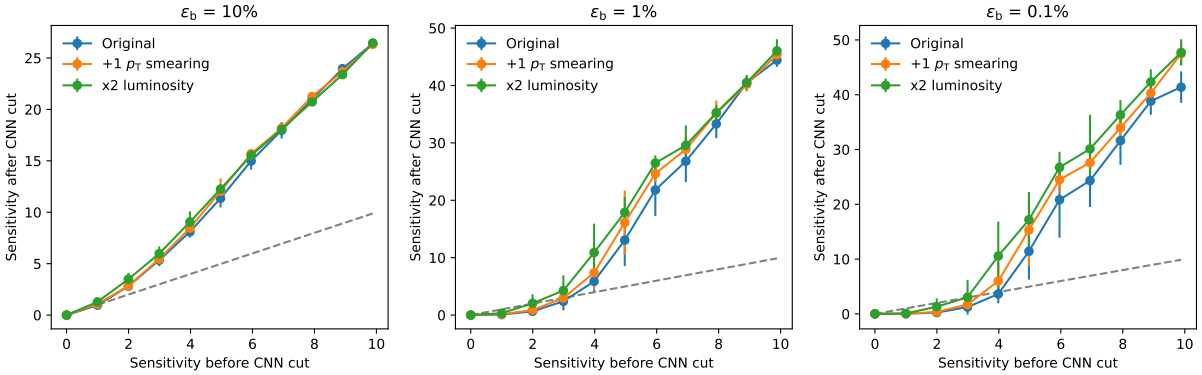
(b) After augmentation

Figure 95: The jet images before and after the p_{T} smearing augmentation.

Figure 96 shows the sensitivity improvement. The p_{T} smearing datasets perform similarly to original datasets. As the background efficiency decreases, the improvement increases. However, because of the large standard deviation, there is no significant difference between original and augmented datasets. If we consider the same sample size, augmented samples have room for improvement.



(a) Resolution: 75×75



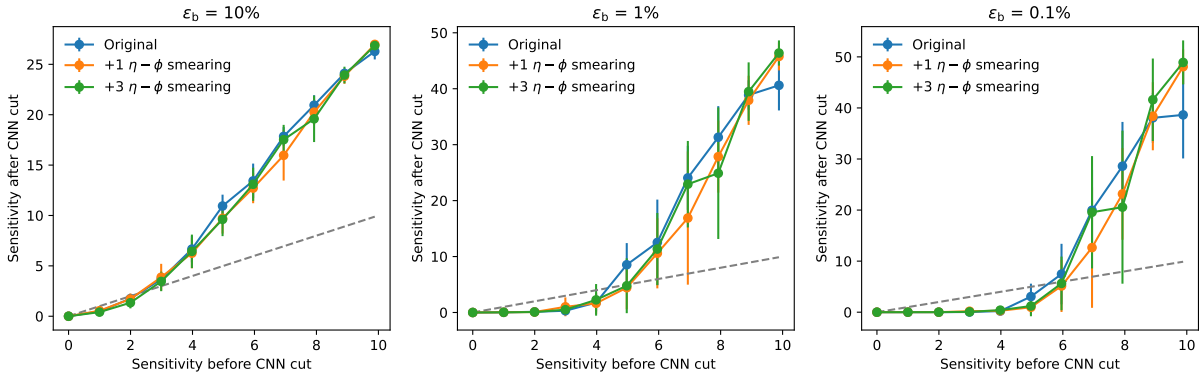
(b) Resolution: 25×25

Figure 96: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

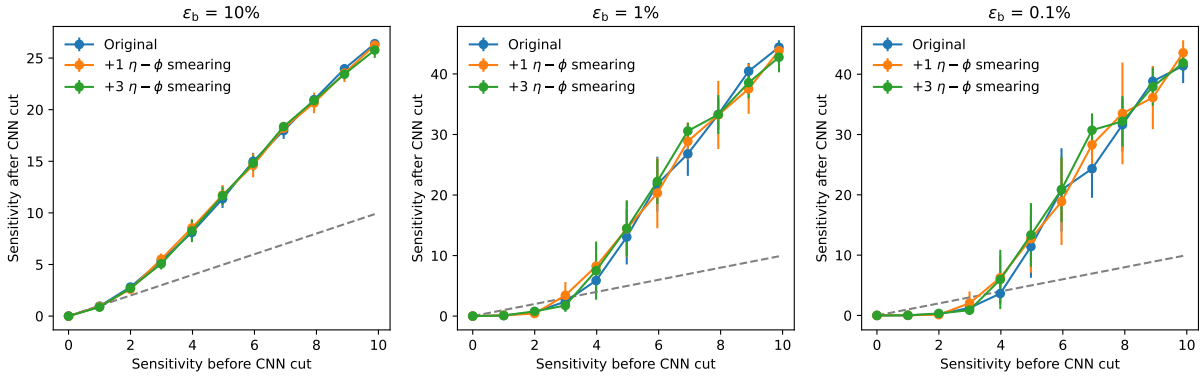
4.33 More augmented sample

Figure 32 shows similar performance for original and $\eta - \phi$ smearing datasets. Figure 96 shows a little improvement for p_{T} smearing samples. However, there is no significant difference because of the large standard deviation. We enlarge the dataset size to ensure whether augmented samples can improve the training.

Figure 97 shows the sensitivity improvement with larger $\eta - \phi$ smearing datasets. Even though we enlarge the training sample size, the results are similar. The $\eta - \phi$ augmentation seems not to improve the training performance.



(a) Resolution: 75×75

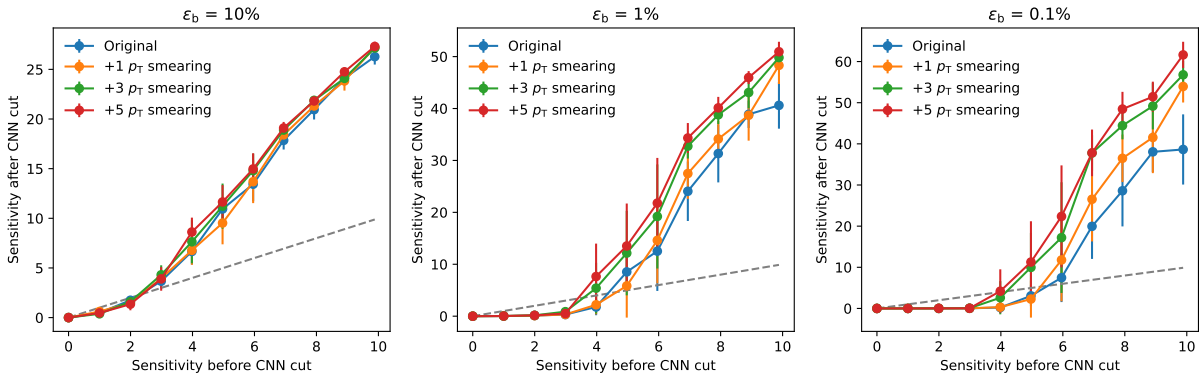


(b) Resolution: 25×25

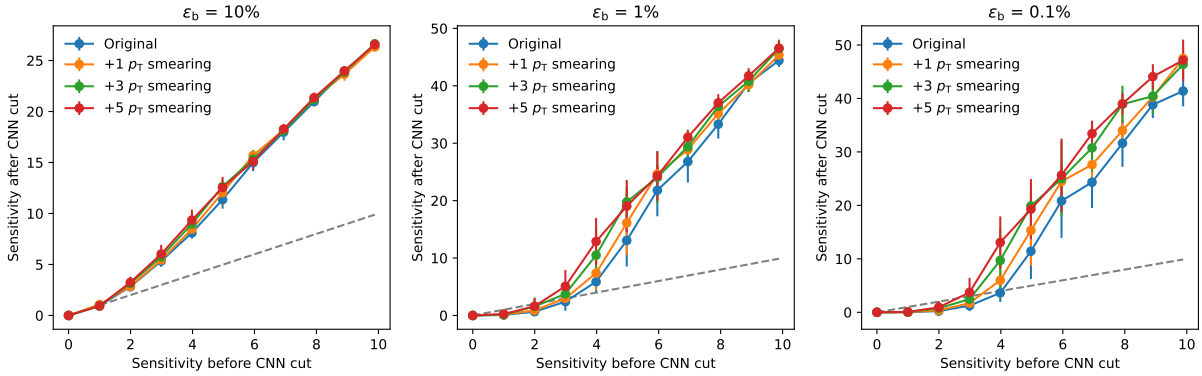
Figure 97: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

Figure 98 shows the sensitivity improvement with larger p_{T} smearing datasets. For the lower background efficiency case, the sensitivity improvement is more significant. As the sample size increases, the training performance becomes better. For $\varepsilon_{\mathrm{b}} = 1\%, 0.1\%$, +3 and +5 p_{T} smearing datasets have lower training thresholds and larger improvement at the high

sensitivity region. The p_{T} smearing can improve the training performance.



(a) Resolution: 75×75



(b) Resolution: 25×25

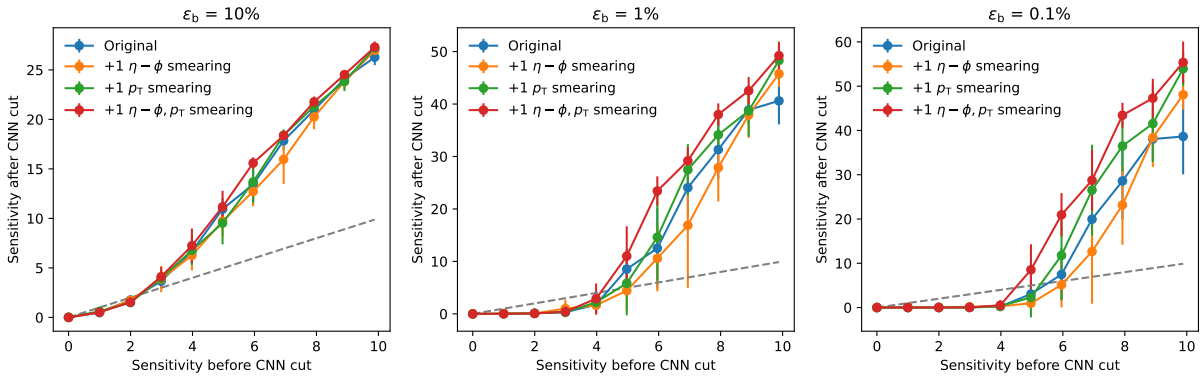
Figure 98: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.34 Combine $\eta - \phi$ and p_{T} smearing

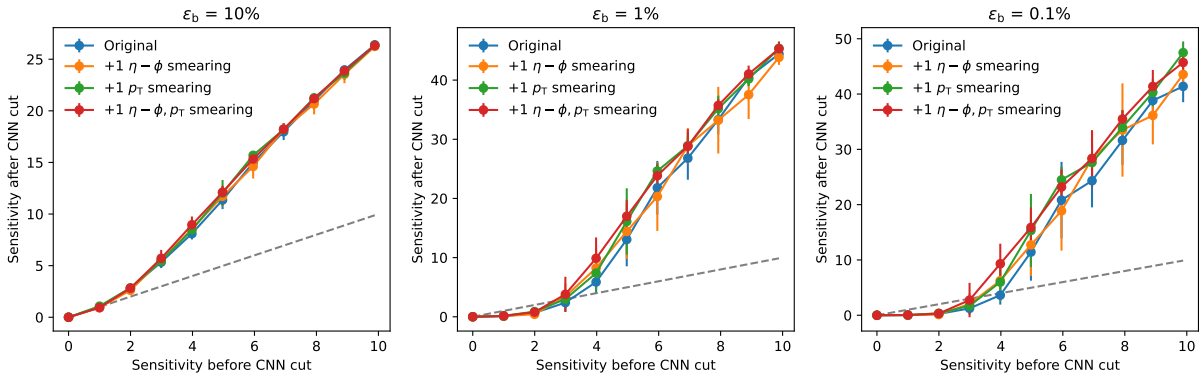
We combine the $\eta - \phi$ and p_{T} smearing, to investigate whether the combining augmentation can further improve training results. We apply $\eta - \phi$ and p_{T} smearing on jet images at the same time for combining augmentation.

Figure 99 shows the sensitivity improvement. The $\eta - \phi$ and p_{T} smearing has the best performance. For the lower background efficiency case, the sensitivity improvement is more notable. There is a significant improvement for the resolution 75×75 with $\varepsilon_{\mathrm{b}} = 0.1\%$. For the resolution 25×25, the performance is similar for p_{T} smearing and combining smearing. However, if we enlarge the training sample size, the difference between p_{T} smearing and

combining smearing would disappear. There is no difference between these two augmentation approaches in Figure 100.



(a) Resolution: 75×75



(b) Resolution: 25×25

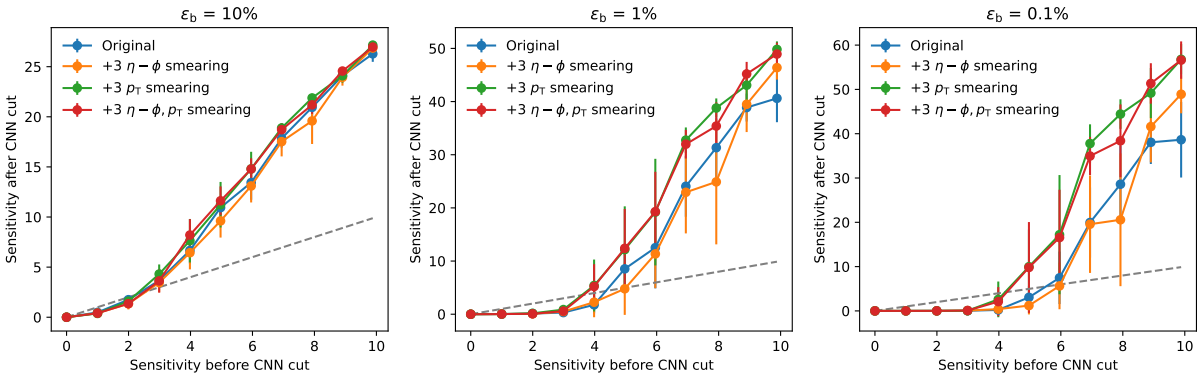
Figure 99: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.35 Jet rotation

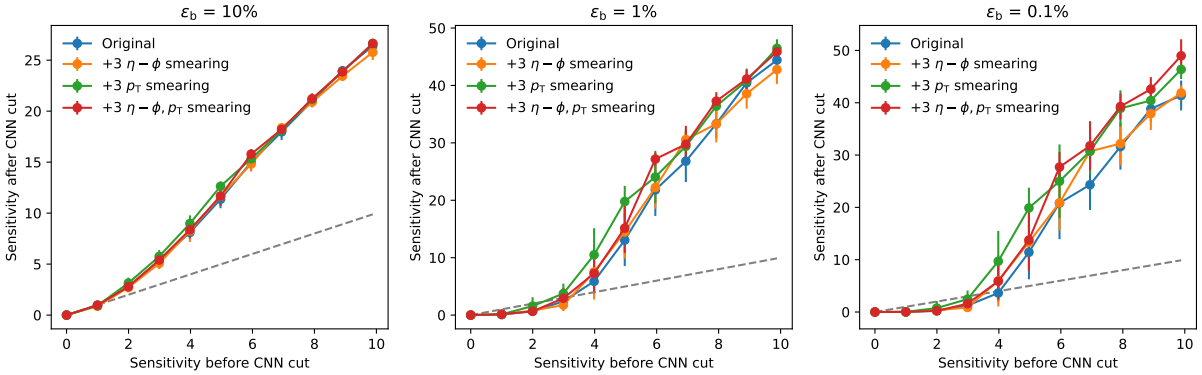
One another augmentation approach is the jet rotation. This method rotates each jet with a random angle to enlarge the diversity of training datasets.

The data process flow is different from Section 4.29. The details are listed below:

1. Utilize the same steps 1 to 3 in Section 4.29.
2. Apply preprocessing on `HDF5` training datasets first.
3. Rotate each jet in each event with a random angle. The rotation angle is uniformly sampled from $[-\pi, \pi]$.



(a) Resolution: 75×75



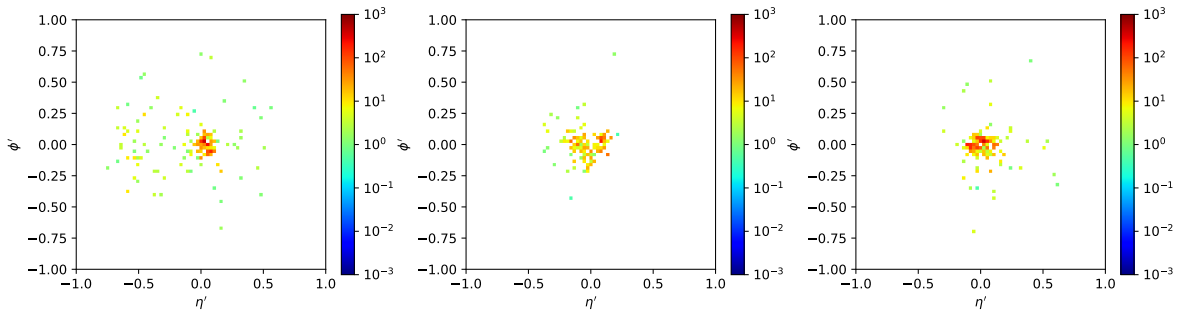
(b) Resolution: 25×25

Figure 100: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

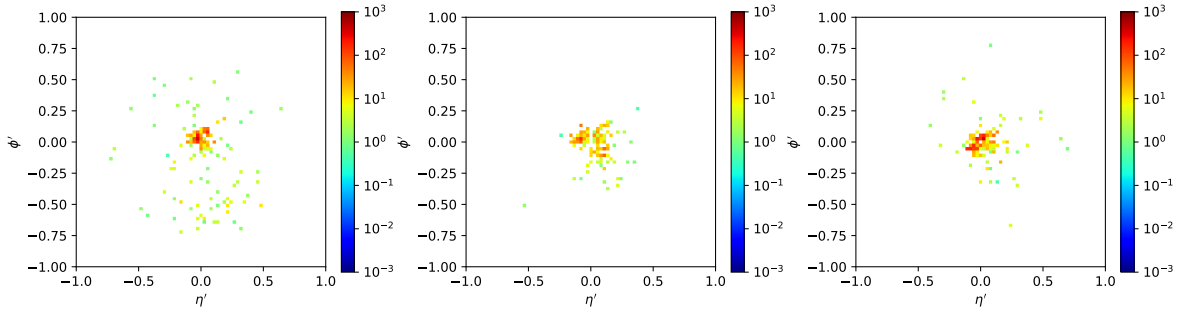
4. Pixelate jets to construct jet images.

5. The validation and testing datasets are the same as in the original case.

Figure 101 is the jet image before and after the jet rotation. These jet images differ by a rotation angle.



(a) Before jet rotation



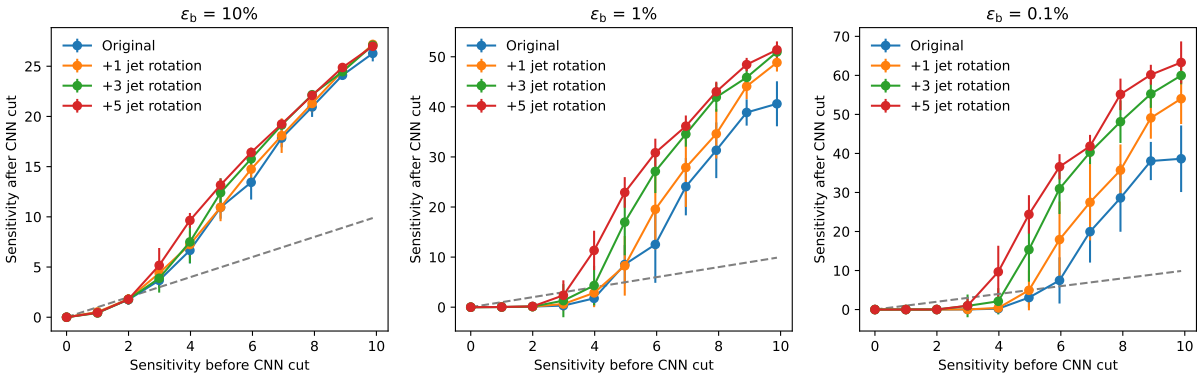
(b) After jet rotation

Figure 101: The jet images before and after the jet rotation augmentation.

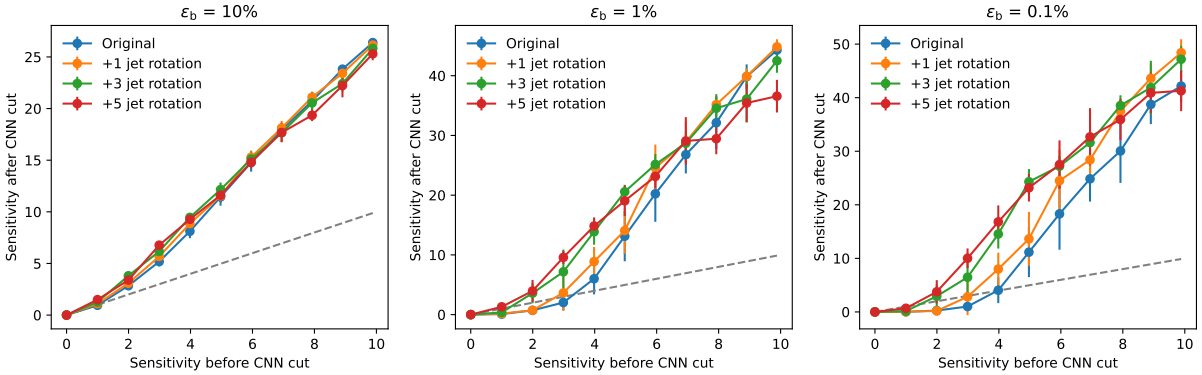
Figure 102 shows the sensitivity improvement with jet rotation datasets. For the lower background efficiency case, the sensitivity improvement is more significant. As the sample size increases, the training performance becomes better. For $\varepsilon_{\mathrm{b}} = 0.1\%$, even if we consider +1 jet rotation datasets, it has lower training thresholds and larger improvement at the high sensitivity region than the original datasets. As the sample size increases, the training performance becomes better. It seems that the jet rotation also can improve the training performance.

4.36 Augmented sample size

To investigate the training performance across different sample sizes, we generate more augmented samples and examine at which point the performance is saturated.



(a) Resolution: 75×75

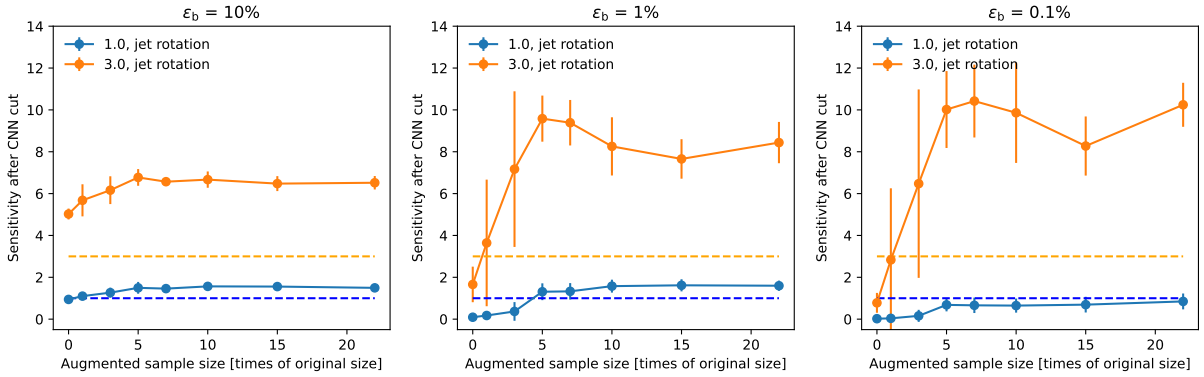


(b) Resolution: 25×25

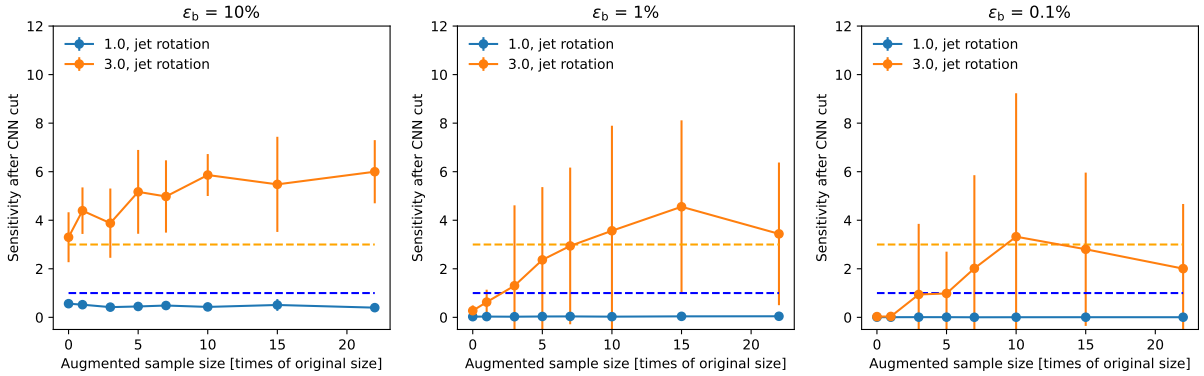
Figure 102: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.36.1 Jet rotation

Figure 103 shows the sensitivity improvement with jet rotation datasets. For resolution 25×25, the training performance is saturated around +5 jet rotation. For resolution 75×75, the training performance is saturated around +15 jet rotation. The higher-resolution model needs more augmented samples to reach the saturation point.



(a) Resolution: 25×25

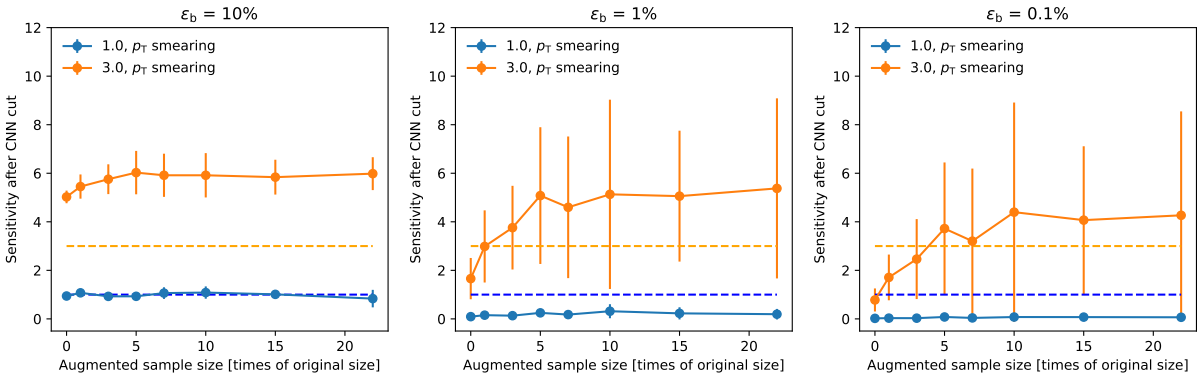


(b) Resolution: 75×75

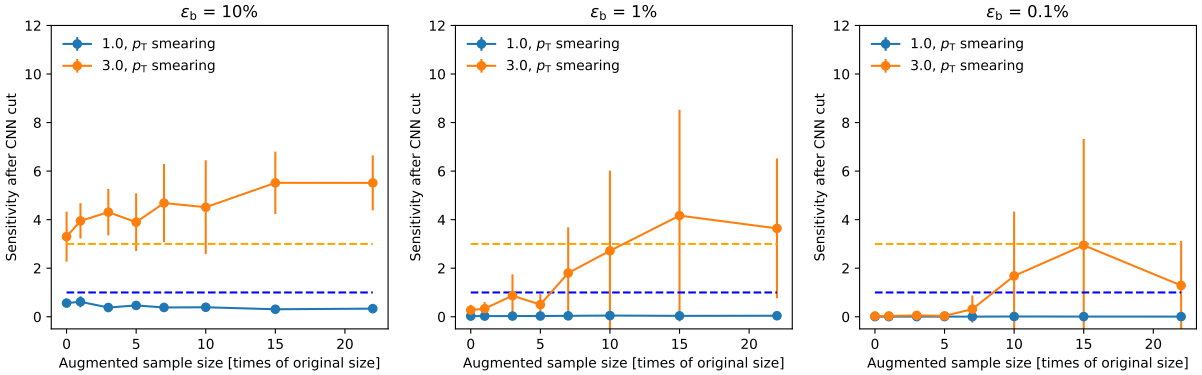
Figure 103: The sensitivities after the CWoLa CNN selection. Here, 1.0 and 3.0 are the sensitivities before selection. The dashed lines are the sensitivities before CNN selection. The error bar is the standard deviation of 10 times training.

4.36.2 p_T smearing

Figure 104 shows the sensitivity improvement with p_T smearing datasets. For resolution 25×25, the training performance is saturated around +5 p_T smearing. For resolution 75×75, the training performance is saturated around +15 p_T smearing.



(a) Resolution: 25×25

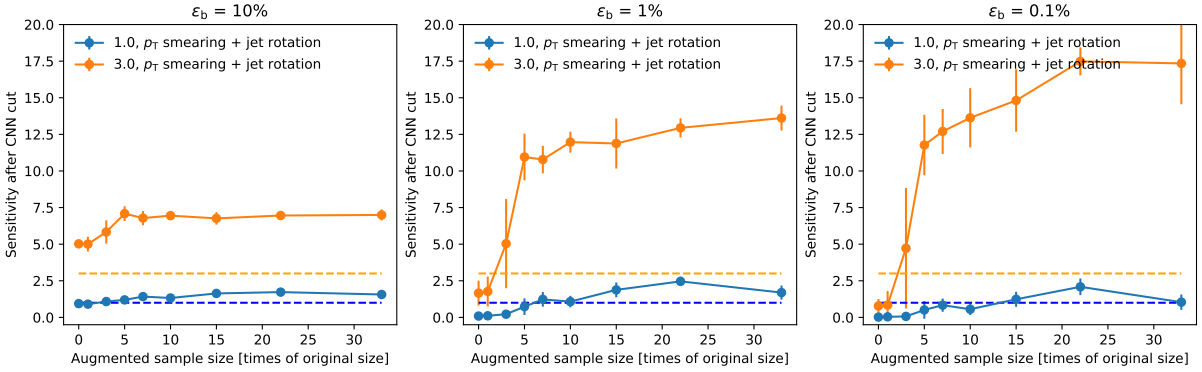


(b) Resolution: 75×75

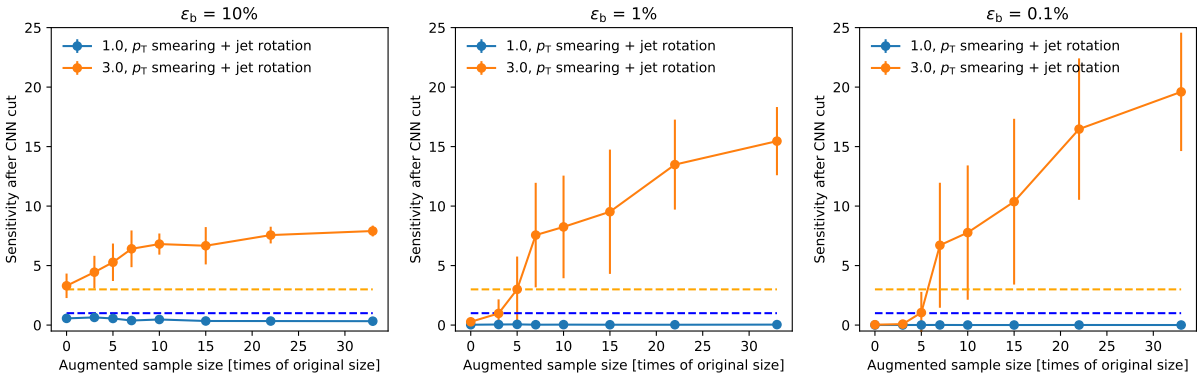
Figure 104: The sensitivities after the CWoLa CNN selection. Here, 1.0 and 3.0 are the sensitivities before selection. The dashed lines are the sensitivities before CNN selection. The error bar is the standard deviation of 10 times training.

4.36.3 p_{T} smearing + Jet rotation

Figure 105 shows the sensitivity improvement of "p_{T} smearing + jet rotation" datasets. For resolution 25×25 with $\varepsilon_{\mathrm{b}} = 10\%$, the sensitivity improvement is saturated around +5 times the augmented size. For lower background efficiency, the performance can still be improved after +5 times augmentation and is saturated around +20 times augmentation. For resolution 75×75, the training performance increases when we use larger datasets. Even if we use +33 times augmented datasets, there seems to be room for improvement for lower background efficiency.



(a) Resolution: 25×25



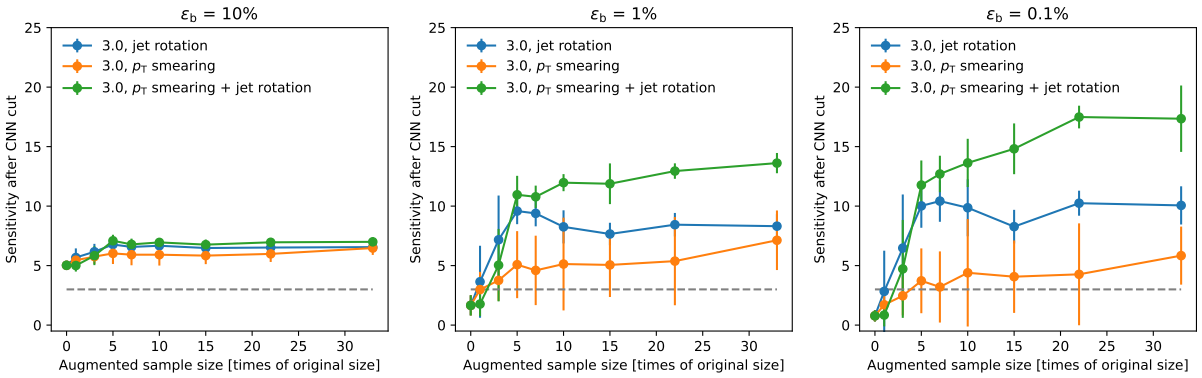
(b) Resolution: 75×75

Figure 105: The sensitivities after the CWoLa CNN selection. Here, 1.0 and 3.0 are the sensitivities before selection. The dashed lines are the sensitivities before CNN selection. The error bar is the standard deviation of 10 times training.

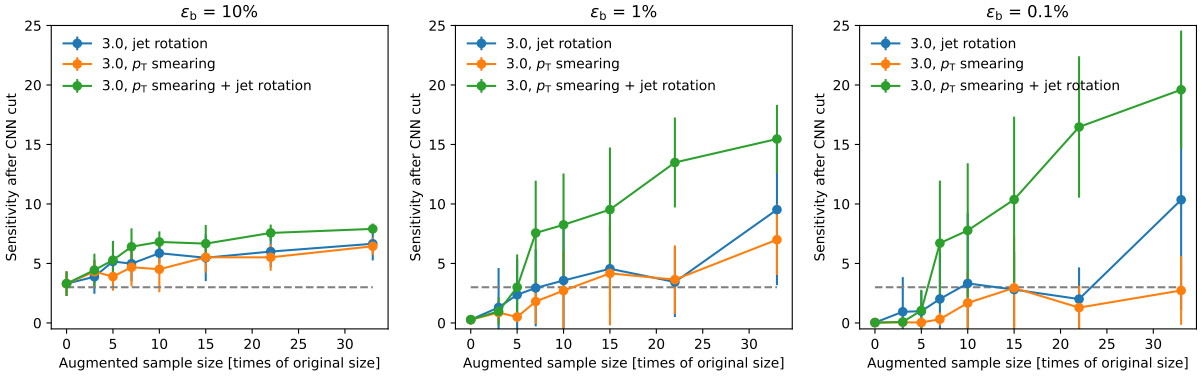
4.36.4 Summary

Figure 106 shows the sensitivity improvement with different augmentation approaches. The "p_{T} smearing + jet rotation" performs best. For resolution 75×75, the combining

method performs much better than other methods.



(a) Resolution: 25×25



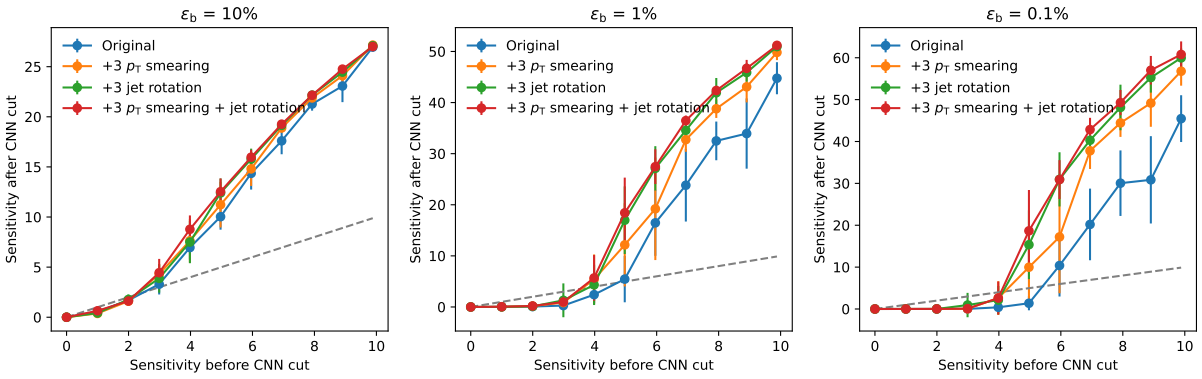
(b) Resolution: 75×75

Figure 106: The sensitivities after the CWoLa CNN selection. Here, 3.0 is the sensitivity before selection. The dashed line is the sensitivity before CNN selection. The error bar is the standard deviation of 10 times training.

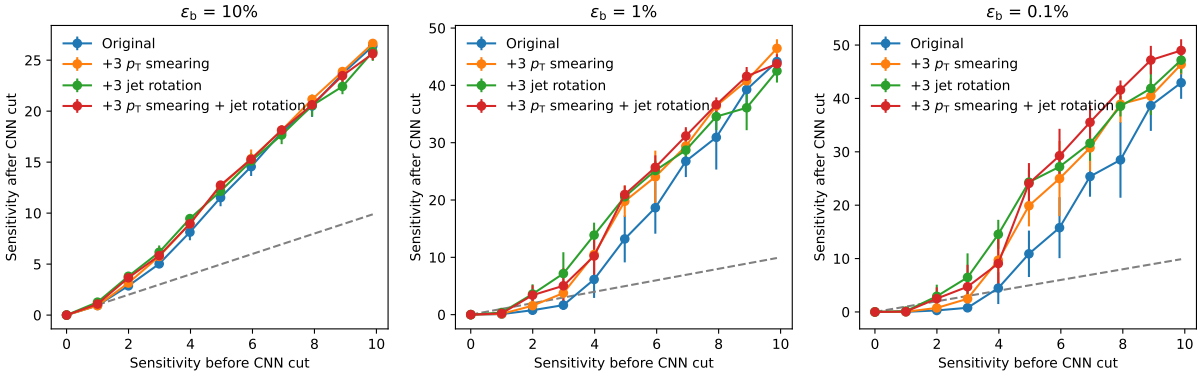
4.37 Combine p_T smearing and jet rotation

We combine the p_T smearing and jet rotation, to investigate whether the combining augmentation can further improve training results. We apply p_T smearing first, then apply jet rotation, because the p_T-weighted center is needed for jet rotation.

Figure 107 shows the sensitivity improvement. All augmentation approaches can improve the training results. The jet rotation is slightly better than the p_T smearing. There seems to be no difference between the "p_T smearing" and "p_T smearing + jet rotation".



(a) Resolution: 75×75

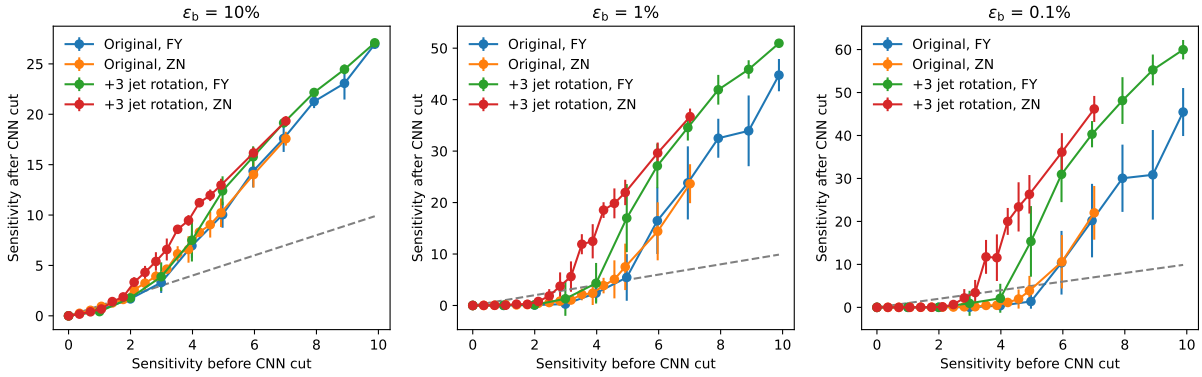


(b) Resolution: 25×25

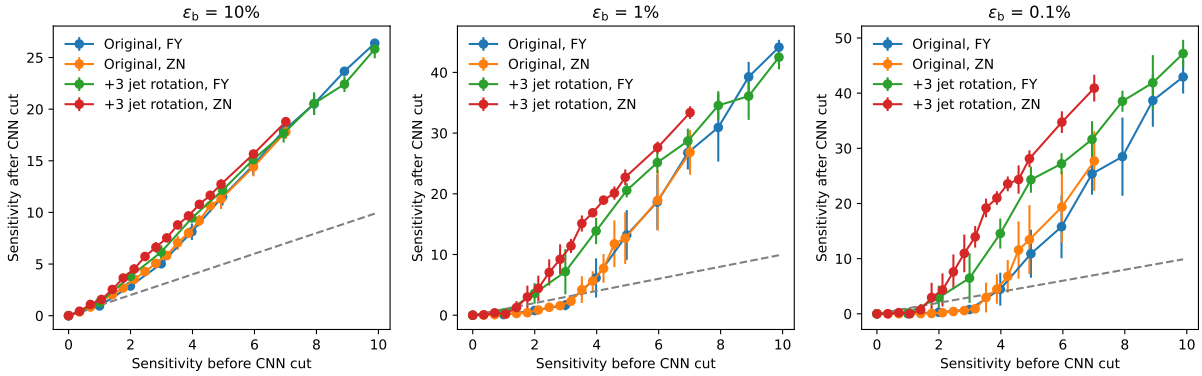
Figure 107: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.38 Compare training results with Zong-En

Figure 108 shows the sensitivity improvement with jet rotation datasets. The sensitivity improvement of FY and ZN are similar for the original datasets. However, ZN's training has better results for jet rotation datasets. From Figure 107, the situation of "p_{T} smearing + jet rotation" augmentation is similar. ZN's training can obtain better results.



(a) Resolution: 75×75



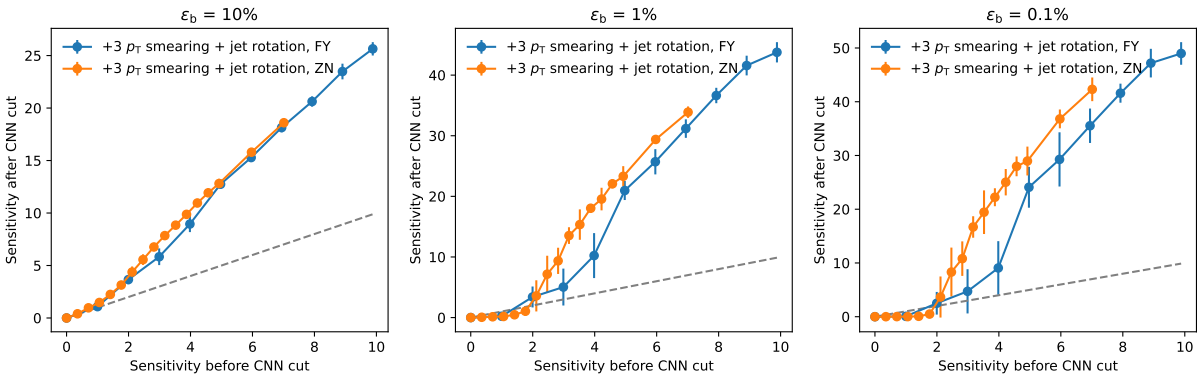
(b) Resolution: 25×25

Figure 108: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

4.39 Fluctuation from different datasets

To explore the fluctuation of different experiments, i.e., the fluctuation of varying training datasets, we prepare a datasets pool and randomly select the events from the pool to make training datasets.

The size of the dataset pool in each category is presented in Table 18. Compared to the



(a) Resolution: 25×25

Figure 109: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

training datasets (Table 17), the background size of the pool is 5 times the training sample size ($\sim$ 20k).

Table 18: The number of events for the dataset pool.

	True label	
Dataset pool	Signal	Background
Signal region	111k	99k
Sideband region	27k	105k

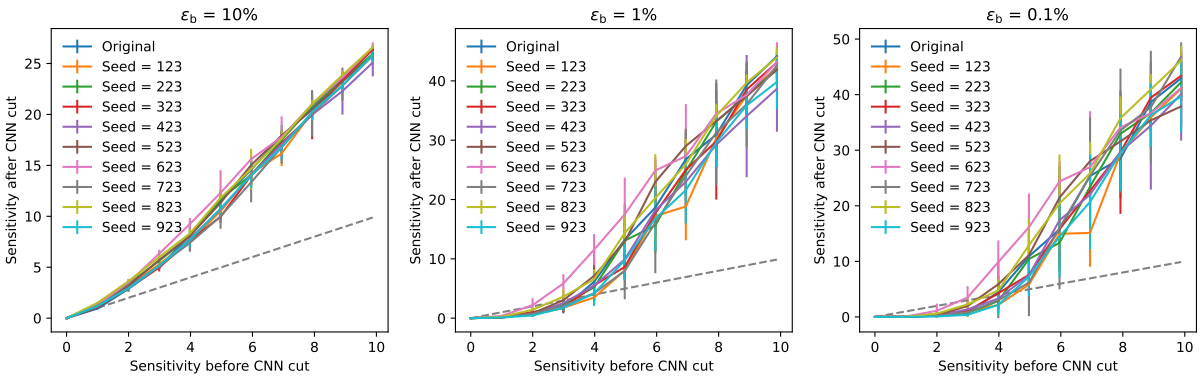
Events are randomly selected from the dataset pool to prepare new training datasets. To reproduce these datasets, the random seed is set by hand.

Figure 110 and Figure 111 show the sensitivity improvement with 10 different datasets. The training could be much different for various datasets. This could explain the differences between ZN's and my training results.

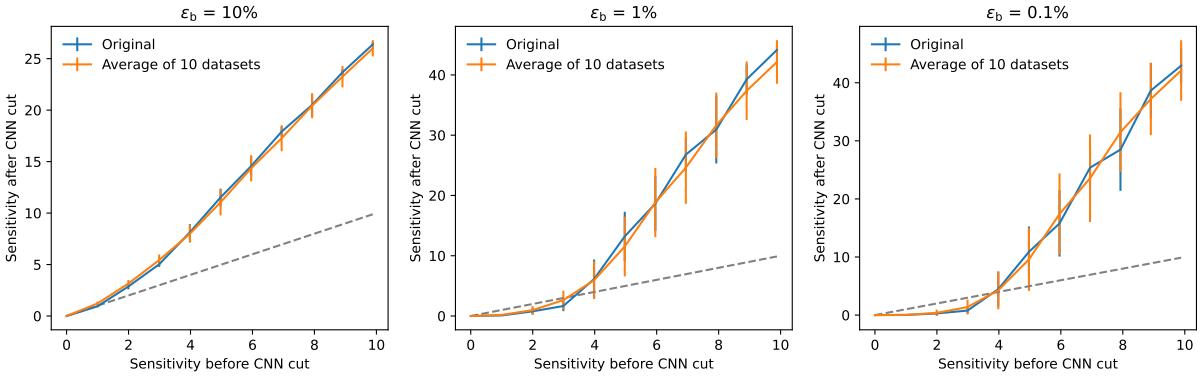
4.40 Hyperparameter optimization

To reduce the fluctuation of the model performance, we tune the hyperparameter setting to optimize the training. For the best hyperparameter set it might make training results more stable. This section uses `Optuna` to do the hyperparameter optimization.

The batch size, learning rate, and some hyperparameters related to the neural network structure are scanned. The optimization range of different hyperparameters is listed in the below

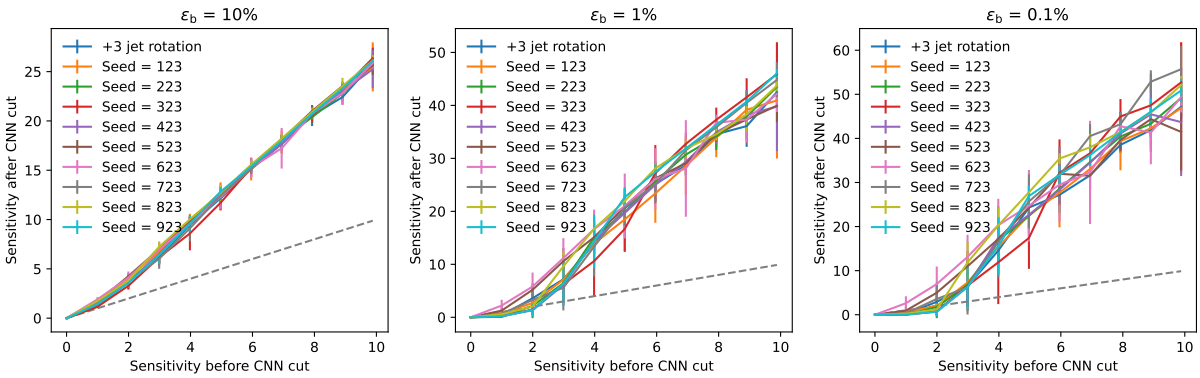


(a) Resolution: 25×25

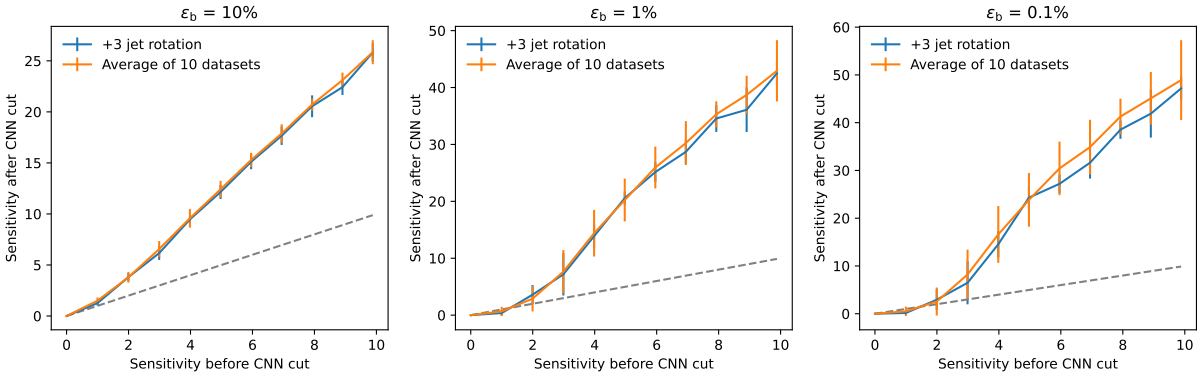


(b) Resolution: 25×25

Figure 110: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The average of 10 datasets is evaluated from 100 times training.



(a) Resolution: 25×25



(b) Resolution: 25×25

Figure 111: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The average of 10 datasets is evaluated from 100 times training.

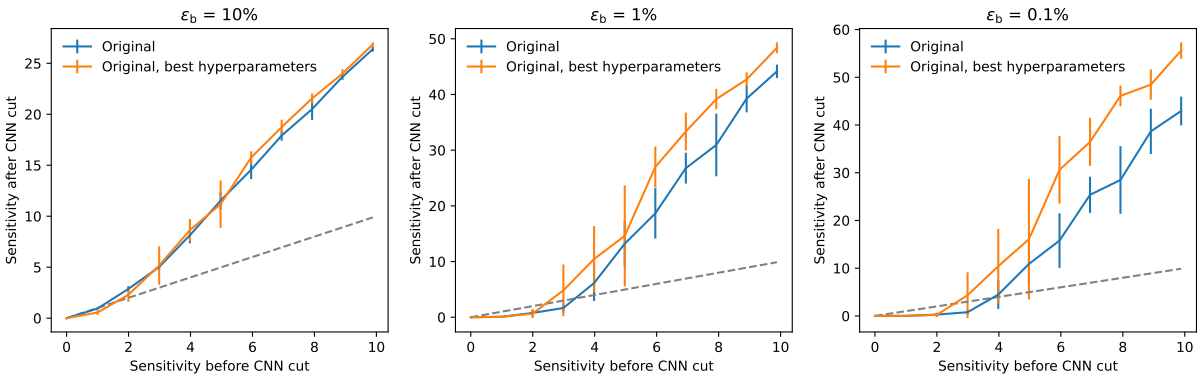
- `BATCH_SIZE`: $[256, 512, 1024, 2048]$
- `learning_rate`: $[10^{-5}, 10^{-2}]$
- `n_CNN_layers_tot`: $[2, 4]$
- `n_CNN_layers_1`: $[1, \texttt{n\_CNN\_layers\_tot}]$
- `n_CNN_filters`: $[16, 32, 64, 128, 256]$
- `CNN_kernel_size`: $[2, 6]$
- `n_dense_layers`: $[1, 5]$
- `dense_hidden_dim`: $[16, 32, 64, 128, 256]$

Test 100 trials and find the parameters set that can minimize the loss value of the testing dataset with true labels.

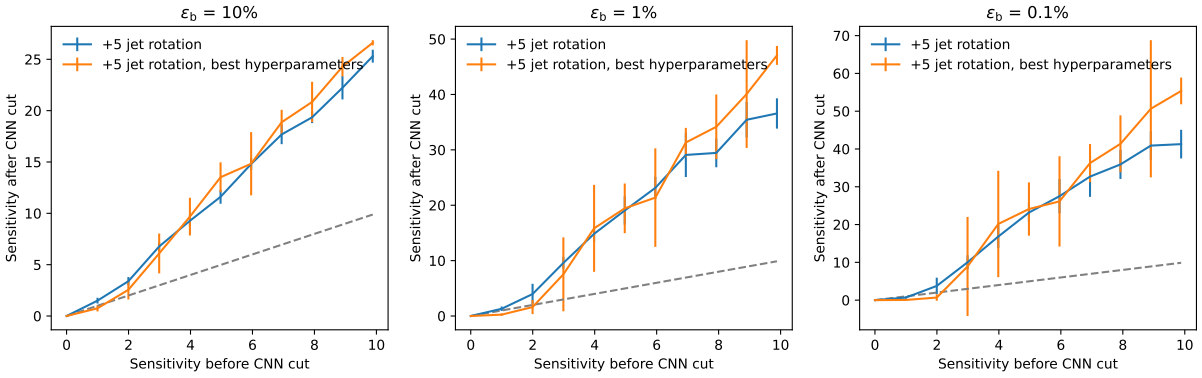
For orignal dataset with $S/\sqrt{B} = 5.0$ and resolution 25×25, the hyperparameter optimization results are listed below

- `BATCH_SIZE`: 256
- `learning_rate`: 2.54×10^{-3}
- `n_CNN_layers_tot`: 2
- `n_CNN_layers_1`: 1
- `n_CNN_filters`: 64
- `CNN_kernel_size`: 6
- `n_dense_layers`: 5
- `dense_hidden_dim`: 128

Test this parameter set on all sensitivities. Figure 112 shows the sensitivity improvement. The training could be improved with this hyperparameter set for original datasets. However, for +5 jet rotation, the performance is similar. The fluctuation of training results could not be reduced for both cases.



(a) Original



(b) Jet rotation: +5

Figure 112: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The average of 10 datasets is evaluated from 100 times training.

4.41 Sculpting effect

To confirm whether the sculpting effect exists for augmented training datasets, we plot the ACC and AUC curves and focus on the performance of no-signal models. Their ACC and AUC values should be close to 0.5 since they are trained on all background datasets.

Figure 113, 114 and 115 are ACC and AUC curves with various augmented approaches. The ACC and AUC values are evaluated from the testing datasets with true labels. We verify that the ACC and AUC values of no-signal models are close to 0.5. No significant difference exists between the original and the augmented cases on the no-signal model. Therefore, the improvement of the augmented dataset does not result from the sculpting effect.

4.42 Jet rotation range

We modify the rotation angle range to investigate how the jet rotation improves the training. For previous training, the range is $\pm 180°$. The small and large ranges are tested. Similarly, the rotation angle is uniformly sampled from the rotation range.

Figure 116 shows the sensitivity improvement of jet rotation datasets with different rotation ranges. For ranges $\pm 1°$ and $\pm 5°$, the sensitivity improvement is close to the original case, since there is no difference in jet images if the rotation angle is tiny. For ranges $\pm 45°$ and $\pm 90°$, their training performance is similar and significantly better than the original case. The dataset with the largest rotation range $\pm 180°$ can achieve the lowest training threshold even at the high sensitivity range the performance is not so good.

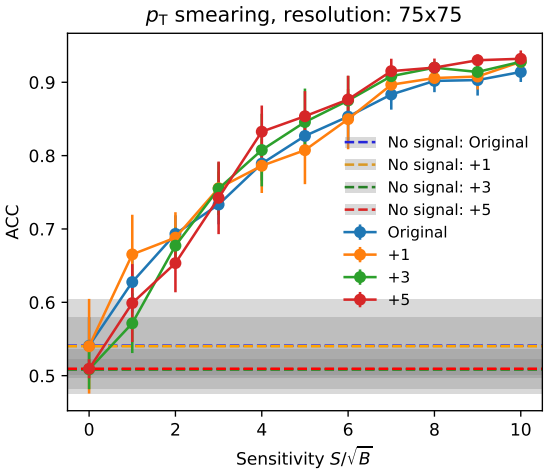
4.42.1 Fixed rotation angles

Instead of using the angle randomly sampled from the rotation range, we specify the rotation angles by hand.

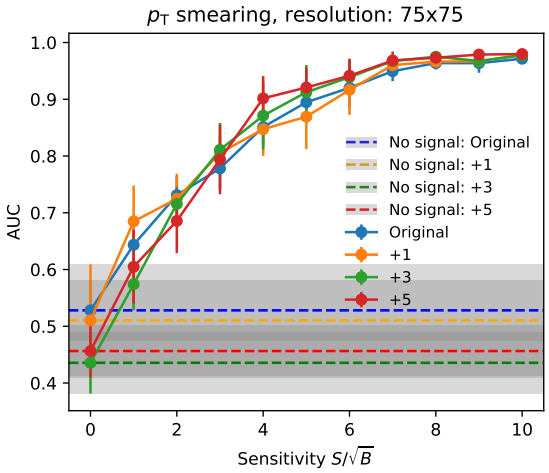
We test $\pm 5°$ rotation. More specifically, the training dataset contains original events, events rotated by $5°$, and events rotated by $-5°$. Therefore, the training sample size is 3 times the original datasets. Figure 117 shows the sensitivity improvement of jet rotation datasets with fixed angles. The sensitivity improvement is close to the original case since there is no difference in jet images if the rotation angle is tiny. Or lower background efficiency, there is a little improvement at the high sensitivity range.

4.43 Re-sampling before training

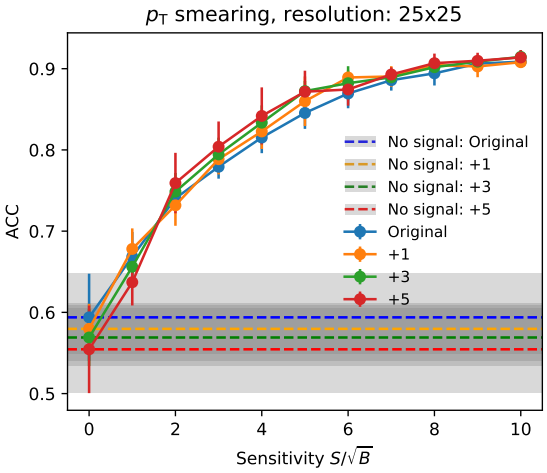
We randomly select events before each training to incorporate the fluctuation of different datasets. The sample process flow is modified as follows to implement the re-sampling:



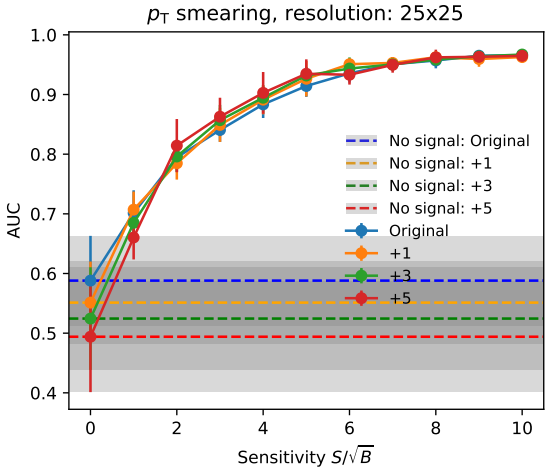
(a) ACC



(b) AUC

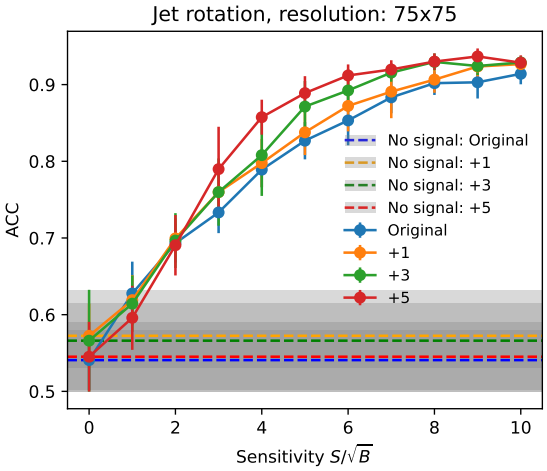


(c) ACC

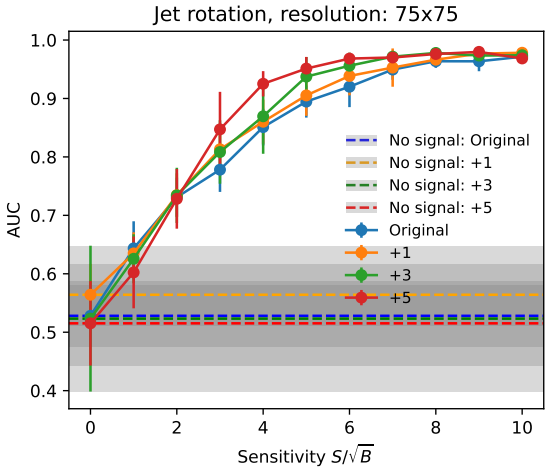


(d) AUC

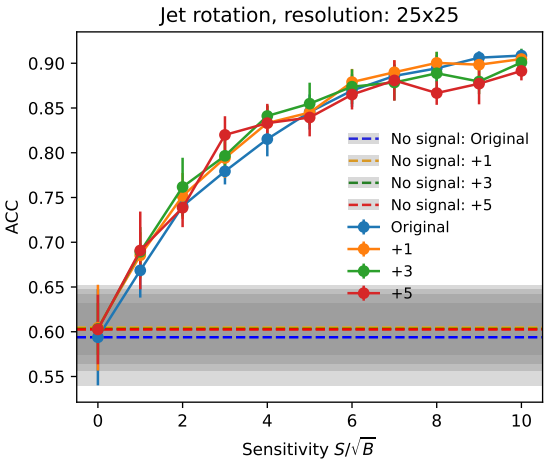
Figure 113: The performance of CWoLa CNN training with p_T smearing datasets. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.



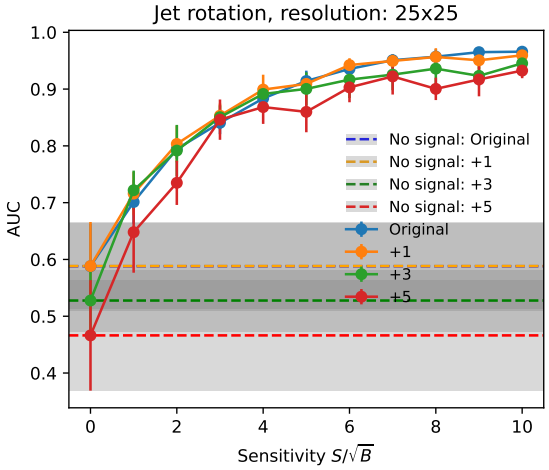
(a) ACC



(b) AUC

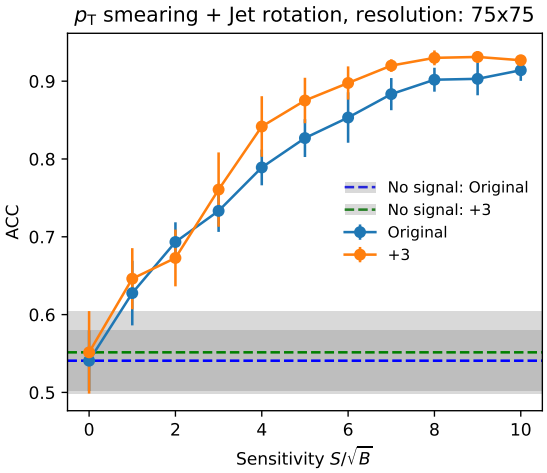


(c) ACC

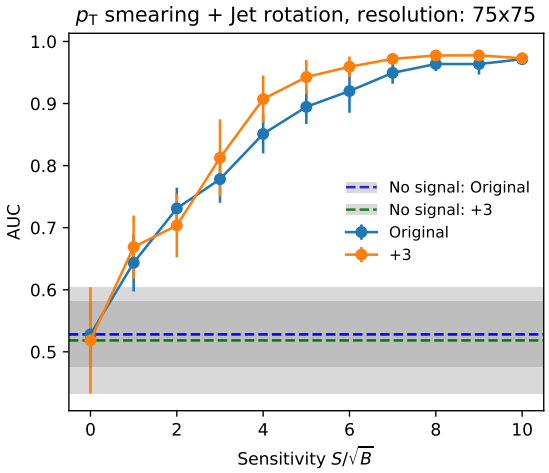


(d) AUC

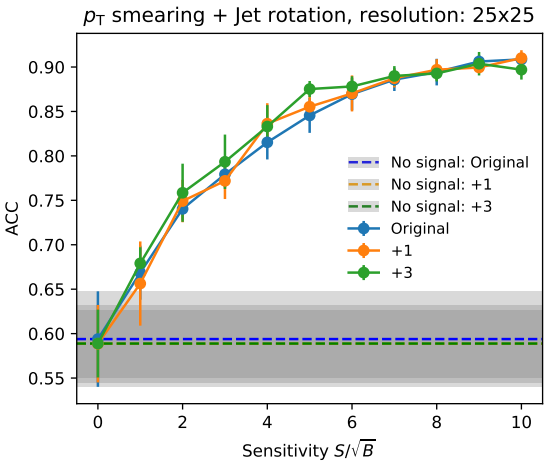
Figure 114: The performance of CWoLa CNN training with jet rotation datasets. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.



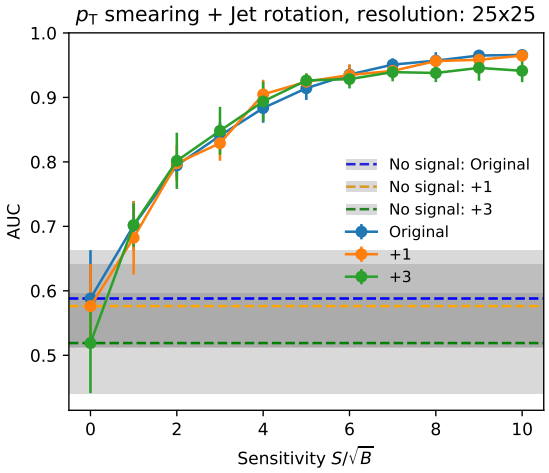
(a) ACC



(b) AUC

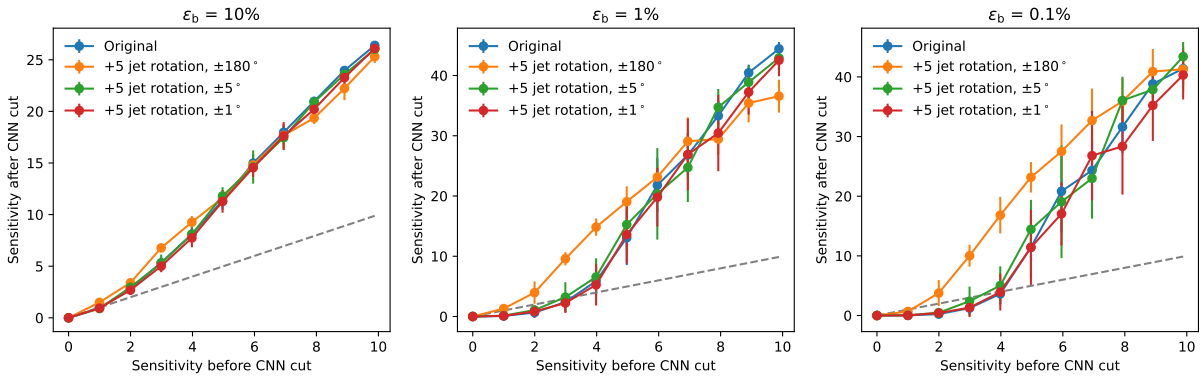


(c) ACC

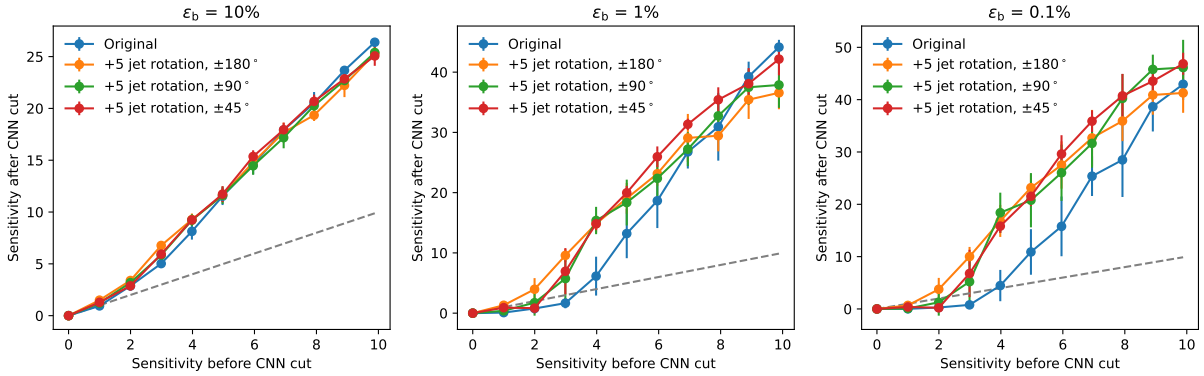


(d) AUC

Figure 115: The performance of CWoLa CNN training with "p_T smearing + jet rotation" datasets. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

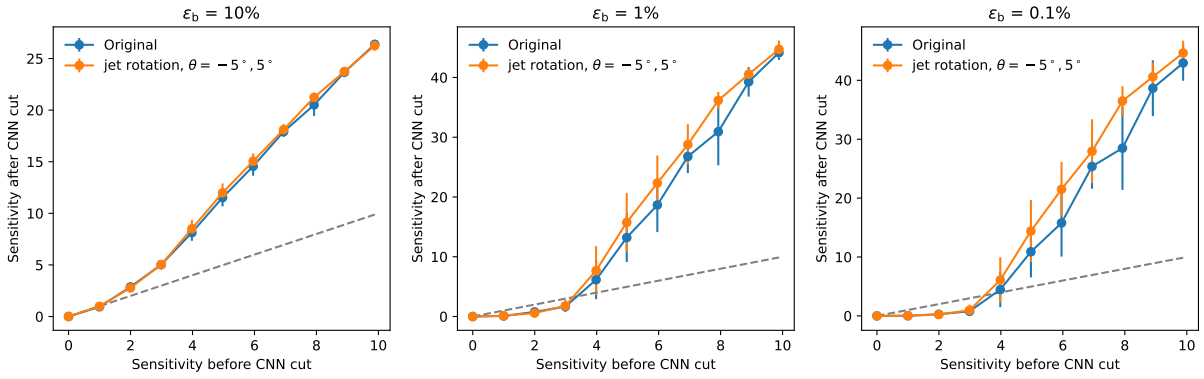


(a) Resolution: 25×25. Rotation range: $\pm 180°, \pm 5°, \pm 1°$



(b) Resolution: 25×25. Rotation range: $\pm 180°, \pm 90°, \pm 45°$

Figure 116: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.



(a) Resolution: 25×25. Rotation angles: $-5°, 5°$

Figure 117: The sensitivities before and after the CWoLa CNN selection. The rotation angle $\theta = -5°, 5°$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.

1. Generate the sample file in `.root` format. Following Section 4.1.

2. Apply the selection cuts described in Section 4.3 and save the event passing the cuts in `HDF5` format. Note that the sideband region is $[4400, 4700] \cup [5500, 5800]$ GeV. The file contains the information listed below

   - The $(p_{\mathrm{T}}, \eta, \phi)$ of leading and sub-leading jet constituents.
   - Total invariant mass m_{jj}.
   - Type of event: 1 for signal, 0 for background.

3. Samples are separated in 4 `HDF5` files: signal in signal region, signal in sideband region, background in signal region, and background in sideband region.

4. Process the sample at `HDF5` level. Preprocess or apply data augmentation. Note that all processing should be done in this step except the pixelization. Results are saved in a directory that contains 4 `HDF5` files. Events with the same index in each directory correspond to the identical original event.

5. Pixelize each event to generate the jet image from `HDF5` data and save in `.npy` file.

6. This step is executed at the head of each training. Following Section 4.5, we can compute the size of datasets. 80% for training set, 20% for validation set. The events with specific indices are chosen to make mixed samples. These indices are obtained from the given random number seed and sensitivity.

   For example, for the "+3 jet rotation" dataset, we select the events from the following directories:

   - `<path-to-origin>/`
   - `<path-to-jet_aug>/01/`
   - `<path-to-jet_aug>/02/`
   - `<path-to-jet_aug>/03/`.

   We can obtain the indices to select events from the given sensitivity and random seed. These indices would repeatedly apply to each directory. Then, the results are "+3 jet rotation" datasets. Note that events with the same index in each directory should correspond to the same original event. Therefore, we do not shuffle events in the data process flow.

Figure 118 shows the sensitivity improvement with the re-sampling process. The "old" curves are the previous training results, where the neural network is trained on the same dataset 10 times. The "new" curves are the re-sampling results, where different datasets are used for each training. The average values are similar. The standard deviation of the new curve is almost greater than the old one. These results are consistent with Section 4.39.

4.44 $\eta - \phi$, p_{T} smearing + Jet rotation

We combine the $\eta-\phi$, p_{T} smearing and jet rotation, to investigate whether the combining augmentation can further improve training results. We apply $\eta - \phi$ and p_{T} smearing first, then apply jet rotation, because the p_{T}-weighted center is needed for jet rotation.

Figure 119 shows the sensitivity improvement. All augmentation approaches can improve the training results. There seems to be no difference between the "p_{T} smearing + jet rotation" and "$\eta-\phi$ p_{T} smearing + jet rotation". These two methods are a little better than the jet rotation.

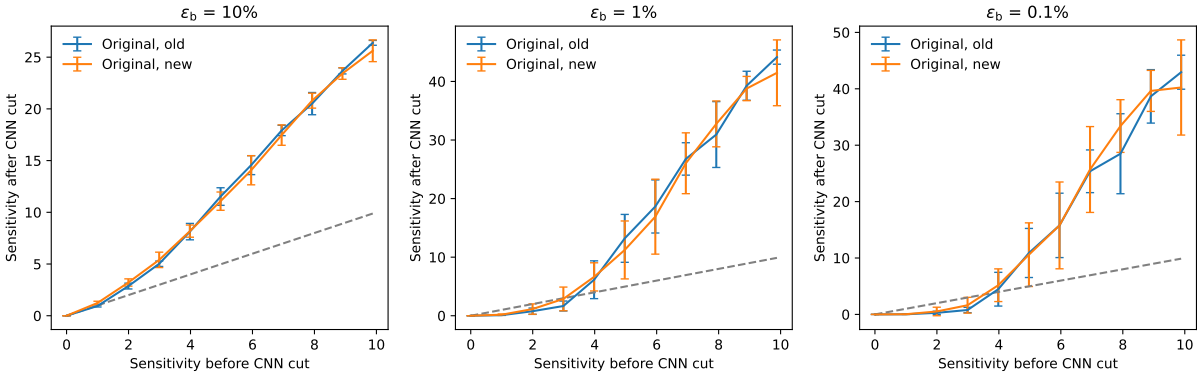
Figure 120 shows the sensitivity improvement with different augmentation approaches. Even though we enlarge the augmented sample size, there is still no difference between the "p_{T} smearing + jet rotation" and "$\eta-\phi$ p_{T} smearing + jet rotation". These two methods are much better than the "jet rotation" augmentation at the large size range. The sensitivity improvement is saturated after +10 times augmentation for all augmentation approaches.

4.45 Check about sculpting

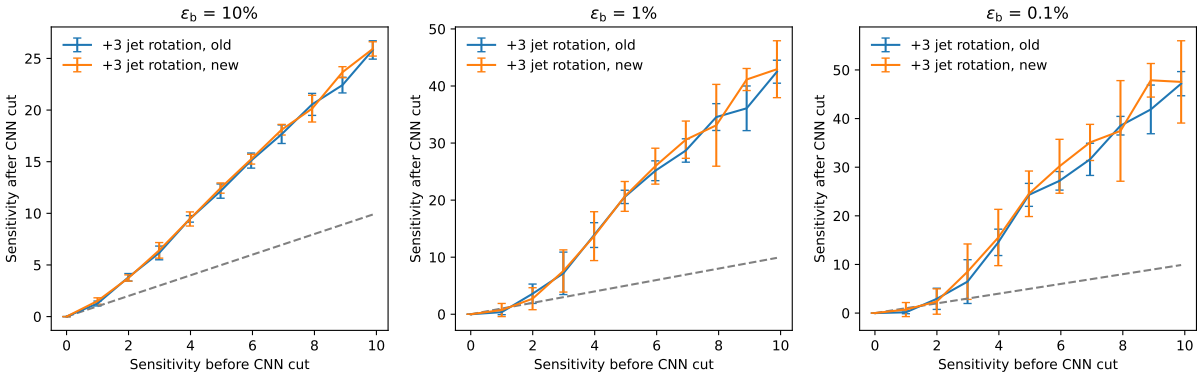
The sculpting effect means the classifier does not learn the difference between the signal and background events and learns the difference between the signal and sideband regions. If a neural network "realizes" what is the signal or background event and has no idea about SR and SB regions, the event score distributions of the SR and SB regions would be the same.

Figure 121 shows the event score distribution of SR and SB regions. We consider original, +3 jet rotation, and +3 p_{T} smearing + jet rotation datasets. All distributions are normalized so that the area of each histogram is 1. The ratio represents the SR value divided by the SB value.

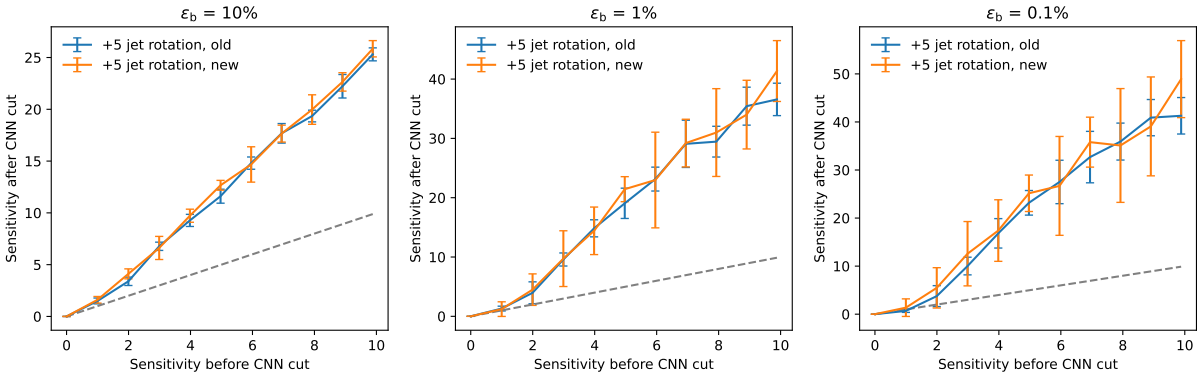
The SR and SB distribution look similar. Due to the small signal sample size, the signal distributions have larger fluctuations. If we examine the distribution closely, the SR distribution appears slightly shifted to the right, but the difference is minimal. Thus, no significant difference exists between SR and SB regions across all cases, indicating no sculpting effect.



(a) Original, Resolution: 25×25

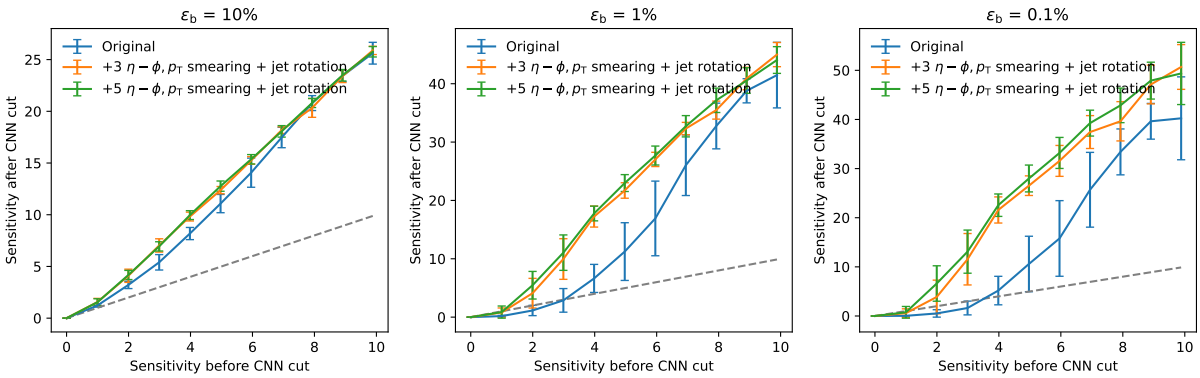


(b) Jet rotation: +3, Resolution: 25×25

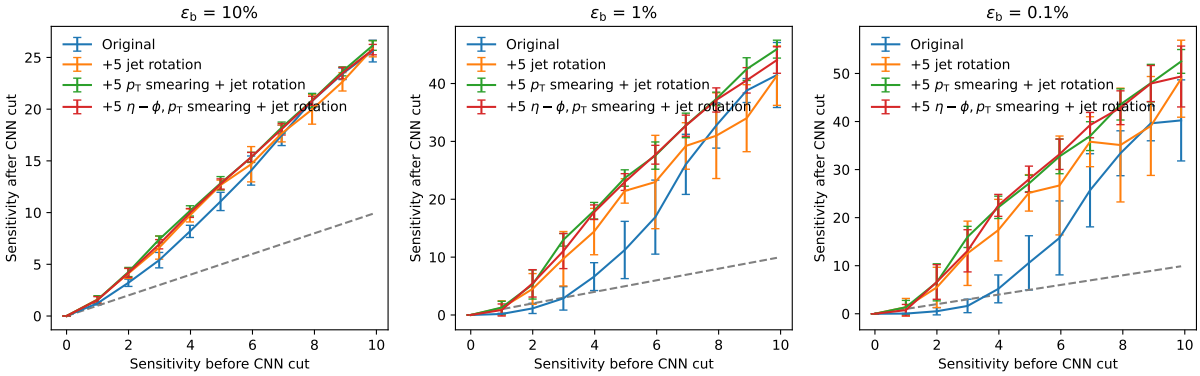


(c) Jet rotation: +5, Resolution: 25×25

Figure 118: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. For new curves, we use different datasets for each training.

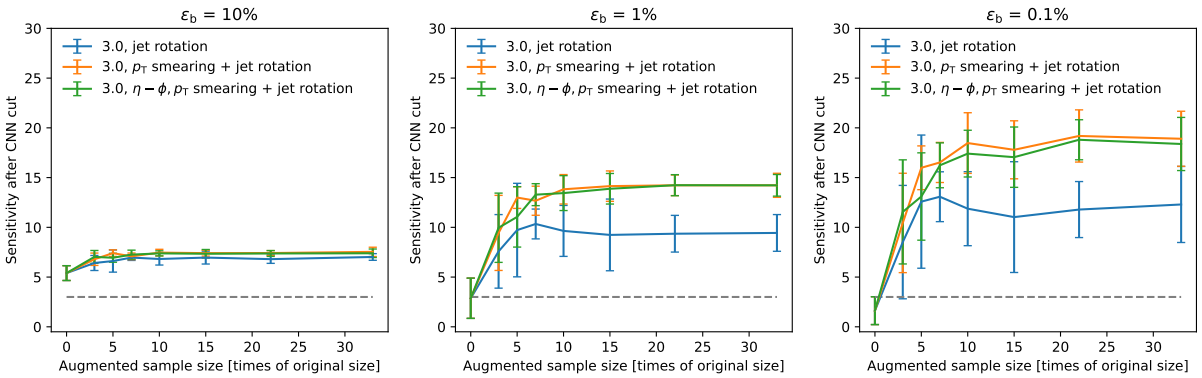


(a) Resolution: 25×25



(b) Resolution: 25×25

Figure 119: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.



(a) Resolution: 25×25

Figure 120: The sensitivities after the CWoLa CNN selection. Here, 3.0 is the sensitivity before selection. The dashed line is the sensitivity before CNN selection. The error bar is the standard deviation of 10 times training.

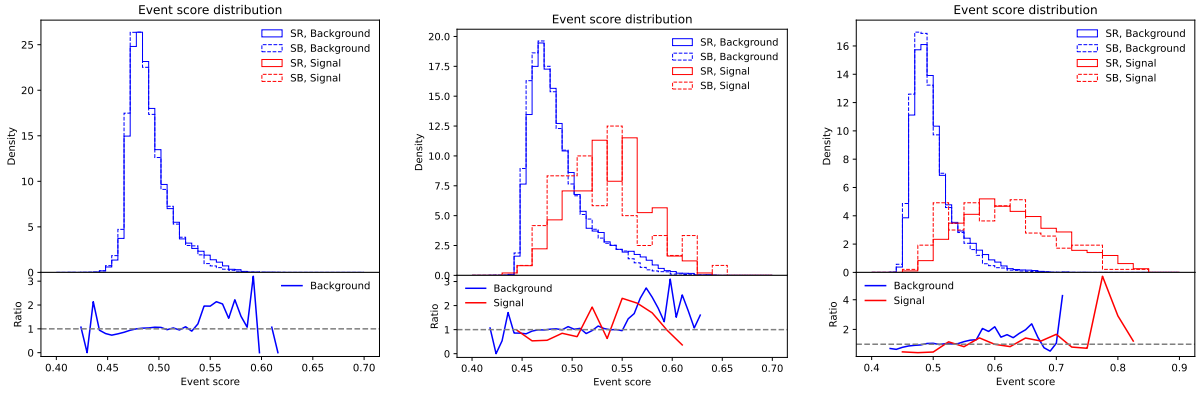
4.46 Threshold value

In previous exercises, we determine threshold values using another testing dataset. This dataset comprises 10k signal and 10k background events in the signal region. In a real application, the CWoLa classifier would be applied back to the training datasets. We would then use the threshold values or sensitivity improvements obtained from the testing sets to compute how many signal and background events remain after applying the cut. We expect the sensitivity improvement from the testing set to be similar to the training set's.

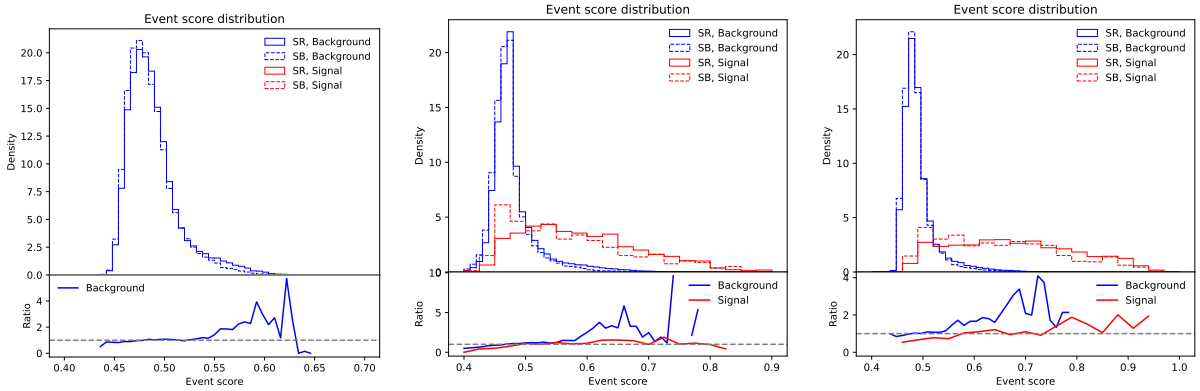
As a sanity check, we should examine whether the training and testing dataset distributions are similar. The threshold values and sensitivity improvements will be identical if they have the same distributions.

Figure 122 shows the event score distributions of the training and testing datasets. We consider original, +3 jet rotation, and +3 p_T smearing + jet rotation datasets. All distributions are normalized so that the area of each histogram is 1. The ratio represents the training value divided by the testing value.

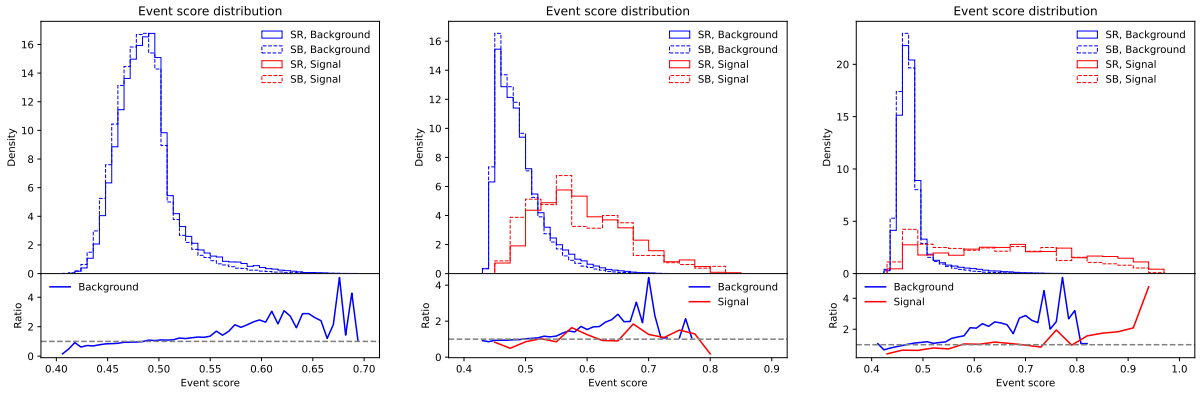
The training and testing distributions look similar. Due to the small sample size, the signal distributions in the training datasets present larger fluctuations. Across all cases, no significant difference exists between training and testing datasets. Thus, we can use another testing set to estimate the sensitivity improvement in the signal region.



(a) Original, $S/\sqrt{B} = 0$ (b) Original, $S/\sqrt{B} = 3$ (c) Original, $S/\sqrt{B} = 7$

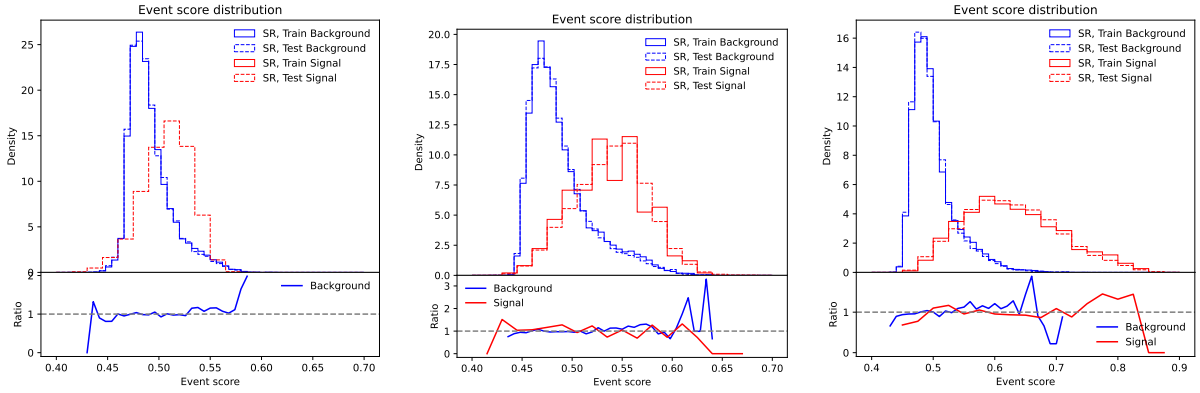


(d) +3 jet rotation, $S/\sqrt{B} = 0$ (e) +3 jet rotation, $S/\sqrt{B} = 3$ (f) +3 jet rotation, $S/\sqrt{B} = 7$



(g) +3 p_{T}-jet, $S/\sqrt{B} = 0$ (h) +3 p_{T}-jet, $S/\sqrt{B} = 3$ (i) +3 p_{T}-jet, $S/\sqrt{B} = 7$

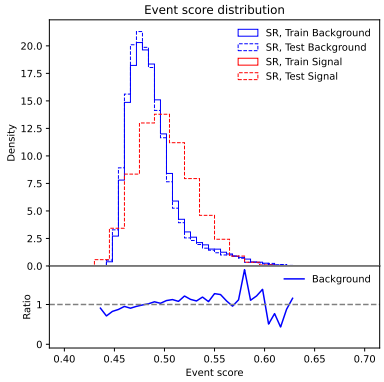
Figure 121: The event score distribution of training samples. Here p_{T}-jet means the p_{T} smearing + jet rotation augmentation.



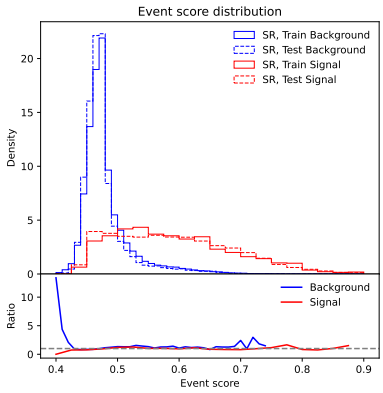
(a) Original, $S/\sqrt{B} = 0$

(b) Original, $S/\sqrt{B} = 3$

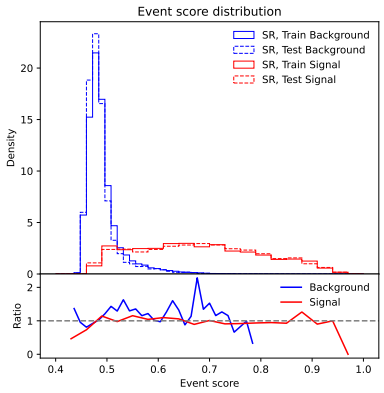
(c) Original, $S/\sqrt{B} = 7$



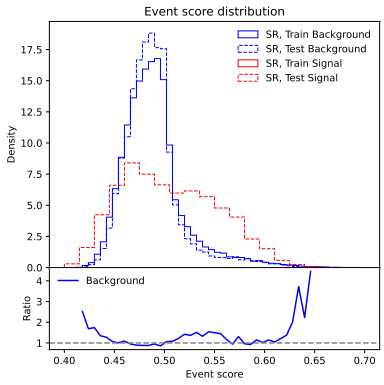
(d) +3 jet rotation, $S/\sqrt{B} = 0$



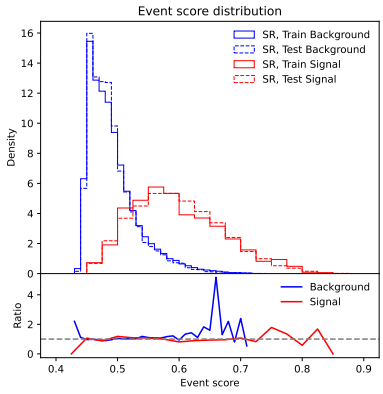
(e) +3 jet rotation, $S/\sqrt{B} = 3$



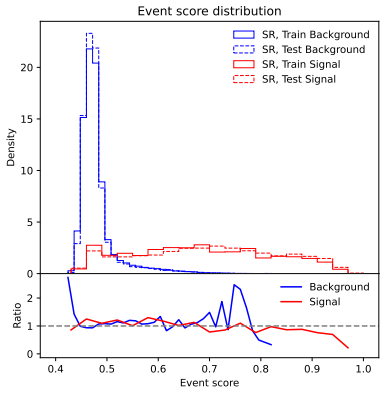
(f) +3 jet rotation, $S/\sqrt{B} = 7$



(g) +3 p_{T}-jet, $S/\sqrt{B} = 0$



(h) +3 p_{T}-jet, $S/\sqrt{B} = 3$



(i) +3 p_{T}-jet, $S/\sqrt{B} = 7$

Figure 122: The event score distribution of training and testing datasets. Here p_{T}-jet means the p_{T} smearing + jet rotation augmentation.

4.47 Sculpting effect

Figure 121 shows that the event score distribution of the SR samples is slightly shifted to the right, which may cause a fake bump in the no-signal case. To determine if the shifting effect is due to overfitting on training datasets, we plot the event score distribution of the testing dataset.

Figure 123 is the event score distribution for the testing dataset, which was not used in training. The ratio represents the SR value divided by the SB value, normalized by the ratio of total events in the SR and SB regions. The grey dashed line, equal to 1, indicates the same shape in both the SR and SB regions. The band is the standard deviation of the ratio, evaluated from the equation:

$$\sigma_{\mathrm{SR/SB}} = \frac{N_{\mathrm{SR}}}{N_{\mathrm{SB}}}\sqrt{\left(\frac{\sigma_{N_{\mathrm{SR}}}}{N_{\mathrm{SR}}}\right)^2 + \left(\frac{\sigma_{N_{\mathrm{SB}}}}{N_{\mathrm{SB}}}\right)^2} = \frac{N_{\mathrm{SR}}}{N_{\mathrm{SB}}}\sqrt{\frac{1}{N_{\mathrm{SR}}} + \frac{1}{N_{\mathrm{SB}}}} \tag{7}$$

where N_{SR} is the number of events in the signal region and $\sigma_{N_{\mathrm{SR}}} = \sqrt{N_{\mathrm{SR}}}$ is the standard deviation of N_{SR}. Similarly, N_{SB} and $\sigma_{N_{\mathrm{SB}}}$ are the corresponding values for the sideband region.

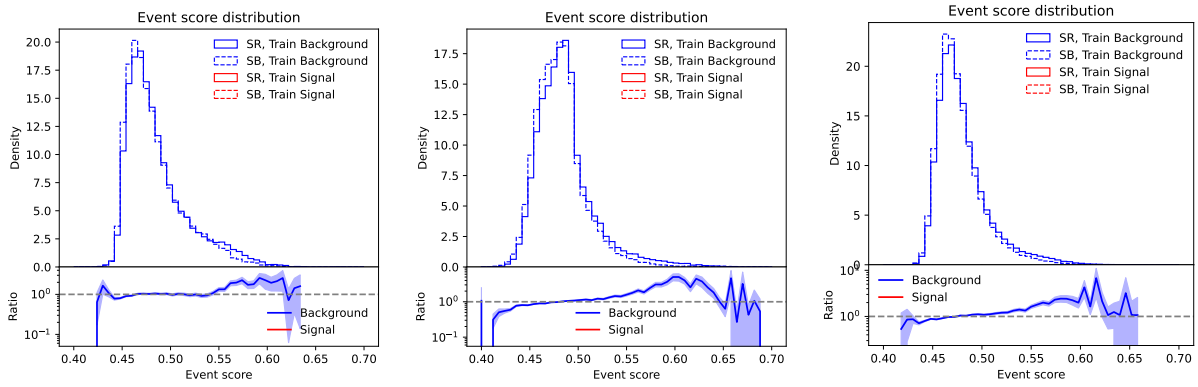
The SR distribution in testing datasets still slightly shifts to the right, which indicates differences between the SR and SB regions.

We evaluate the background passing efficiency in the SR and SB regions to describe the sculpting effect. If the sculpting effect does not exist, we expect the background efficiencies in the SR and SB regions to be the same. However, the sculpting effect makes the background efficiency in the SR higher than in the SB, creating a fake bump. To estimate the sensitivity of this fake bump, we use the following formula

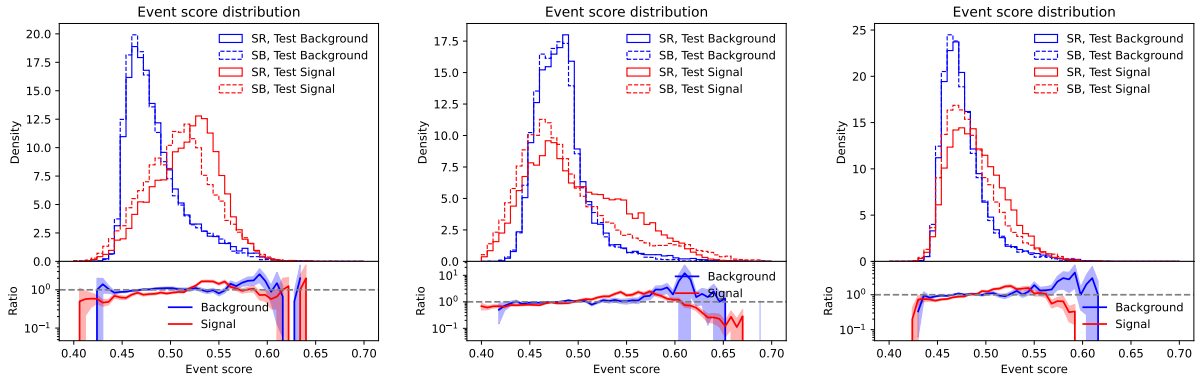
$$\sigma = \frac{B \times (\epsilon_{\mathrm{SR}} - \epsilon_{\mathrm{SB}})}{\sqrt{B\epsilon_{\mathrm{SB}}}}, \quad \mathrm{std}\,(\sigma) = \sqrt{\frac{B}{\epsilon_{\mathrm{SB}}}}\mathrm{std}\,(\epsilon_{\mathrm{SR}}) = \sqrt{\frac{B}{\epsilon_{\mathrm{SB}}}}\sqrt{\frac{\epsilon_{\mathrm{SR}}(1-\epsilon_{\mathrm{SR}})}{N_{\mathrm{SR}}}} \tag{8}$$

where $\epsilon_{\mathrm{SR}}, \epsilon_{\mathrm{SB}}$ are the background efficiencies in SR and SB regions, respectively, B is the number of background events in the signal region, and N_{SR} is the number of background events using to compute ϵ_{SR}.

Table 19 and 20 show the background efficiency in the SR and SB regions. The number of background events $B = 18,922$. The fake sensitivity of training samples is greater than that of testing samples in all cases. This suggests that the sculpting effect partially arises from the overfitting of training samples.



(a) Original, $S/\sqrt{B} = 0$ (b) +3 jet rotation, $S/\sqrt{B} = 0$ (c) +3 p_{T}-jet, $S/\sqrt{B} = 0$



(d) Original, $S/\sqrt{B} = 0$ (e) +3 jet rotation, $S/\sqrt{B} = 0$ (f) +3 p_{T}-jet, $S/\sqrt{B} = 0$

Figure 123: The event score distribution of training samples with resolution 25×25. Here p_{T}-jet means the p_{T} smearing + jet rotation augmentation.

Table 19: The background efficiencies and fake bump sensitivities with resolution 25×25 datasets.

Model	Training sample			Testing sample		
	ϵ_{SB}	ϵ_{SR}	σ	ϵ_{SB}	ϵ_{SR}	σ
Original	10.0%	12.4%	10.3 ± 1.2	10.0%	10.7%	3.1 ± 1.3
	1.00%	2.02%	14.1 ± 1.6	1.00%	1.71%	9.8 ± 1.8
	0.10%	0.22%	5.4 ± 1.7	0.10%	0.06%	-1.7 ± 1.1
+3 Jet rotation	10.0%	14.2%	18.3 ± 0.6	10.0%	11.2%	5.1 ± 1.4
	1.00%	2.68%	23.1 ± 0.9	1.00%	2.16%	16.0 ± 2.0
	0.10%	0.11%	0.5 ± 0.6	0.10%	0.03%	-3.0 ± 0.8
+3 p_{T}-jet	10.0%	13.2%	13.9 ± 0.6	10.0%	11.5%	6.5 ± 1.4
	1.00%	2.11%	15.3 ± 0.8	1.00%	1.63%	8.7 ± 1.7
	0.10%	0.20%	4.4 ± 0.8	0.10%	0.11%	0.4 ± 1.4

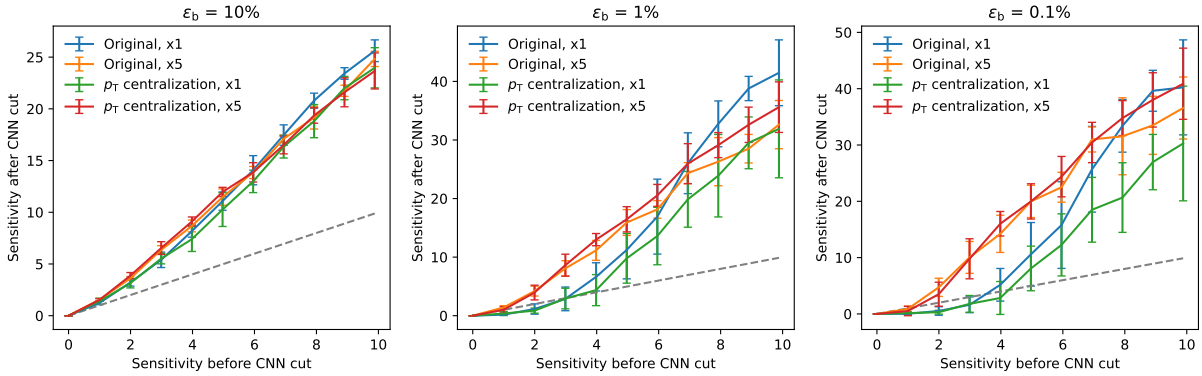
Table 20: The background efficiencies and fake bump sensitivities with resolution 75×75 datasets.

Model	Training sample			Testing sample		
	ϵ_{SB}	ϵ_{SR}	σ	ϵ_{SB}	ϵ_{SR}	σ
Original	10.0%	12.8%	12.2 ± 1.2	10.0%	10.6%	2.7 ± 1.3
	1.00%	2.17%	16.1 ± 1.6	1.00%	1.62%	8.5 ± 1.7
	0.10%	0.52%	18.4 ± 2.5	0.10%	0.16%	2.6 ± 1.7
+3 Jet rotation	10.0%	13.8%	16.4 ± 0.6	10.0%	10.7%	2.9 ± 1.3
	1.00%	2.18%	16.3 ± 0.8	1.00%	1.39%	5.4 ± 1.6
	0.10%	0.40%	13.2 ± 1.1	0.10%	0.20%	4.3 ± 1.9
+3 p_{T}-jet	10.0%	14.1%	18.0 ± 0.6	10.0%	11.3%	5.7 ± 1.4
	1.00%	2.20%	16.6 ± 0.8	1.00%	1.42%	5.8 ± 1.6
	0.10%	0.32%	9.7 ± 1.0	0.10%	0.17%	3.0 ± 1.8

4.48 Remove rotation and flipping preprocessing

To explore how the jet rotation improves the training, we modify the preprocessing steps mentioned in Section 4.4. We remove the rotation and flipping and only keep the p_{T} centralization for training datasets in this section. However, we keep all current preprocessing procedures for testing samples.

Figure 124 shows the sensitivity improvement of different preprocessing procedures. For $\mathcal{L} = 139$ fb^{-1}, the datasets without the rotation and flipping preprocessing perform worse. However, the p_{T} centralization datasets perform similarly or slightly better for $\mathcal{L} = 139 \times 5$ fb^{-1}. If we enlarge the training sample size by increasing the luminosity, the learning threshold would become lower. The "p_{T} centralization with 5 times luminosity" can perform better than the "original with 1 times luminosity" at the low sensitivity region, even if we only apply the network on all preprocessing samples. This gives a reason why the jet rotation can improve performance.



(a) Resolution: 25×25

Figure 124: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training. Here $\times n$ means the luminosity is equal to $139 \times n$ fb^{-1}.

4.49 Invariant mass distribution

We plot the invariant mass m_{jj} distribution before and after the CNN cut to determine whether the sculpting effect exists. We also plot the background passing rate as a function of invariant mass m_{jj} to investigate the performance of CWoLa CNN on different mass regions.

Figure 125 and 126 show m_{jj} distributions and background passing rates for training and testing datasets. The number of events is normalized by the cross-section. The ε_{b} is

the number of events after the CNN cut divided by the value before the CNN cut. The grey dashed line equals the background passing rate at the sideband region. The band is the standard deviation of the ε_{b}, evaluated from the Equation 7. We prepare another testing dataset to plot the m_{jj} distribution here. These samples all pass the η cut mentioned in Section 4.3.

The m_{jj} distributions look smooth for the $\varepsilon_{\mathrm{SB}} = 10\%$ case. However, there are obvious sculpting effects for lower background efficiency cases. These results are consistent with Table 19. The background passing rate increases with the invariant m_{jj}. CNN cuts the low mass region.

4.50 Background subtraction

Figure 125 and 126 show the sculpting effect. We can not simply use the background passing rate in the sideband region to estimate the background passing rate in the signal region.

We would use the background subtraction method:

1. Prepare the real and simulated datasets. In real applications, the real dataset comes from experiments, which could contain the signal and background events. The simulated dataset comes from simulation, which only contains background events in the signal region.

2. Train a CWoLa classifier with the real dataset.

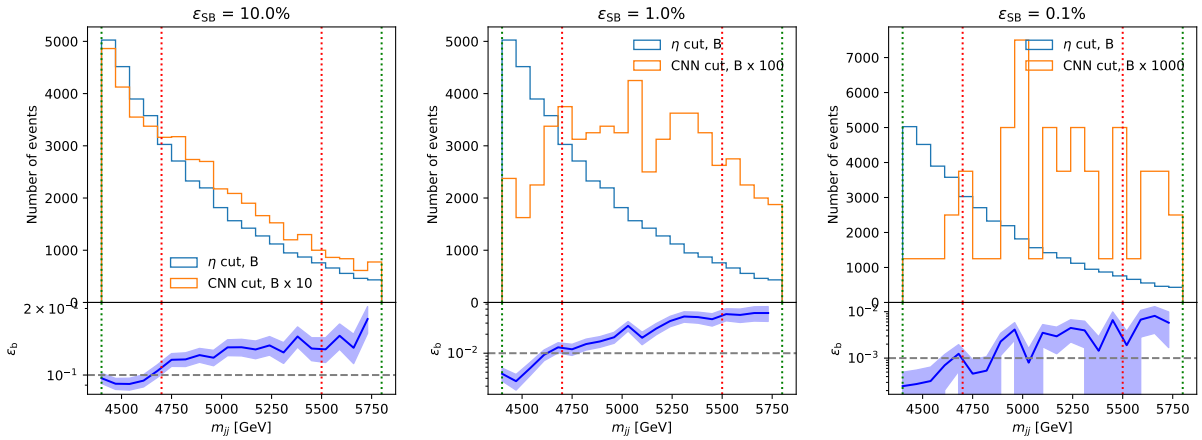
3. Use the simulated dataset to determine the background passing rate and the corresponding threshold.

4. Apply the CWoLa classifier with the same threshold on the real dataset in the signal region. Then we can obtain the number of events passing the CWoLa cut. This value is the sum of the signal and background events.

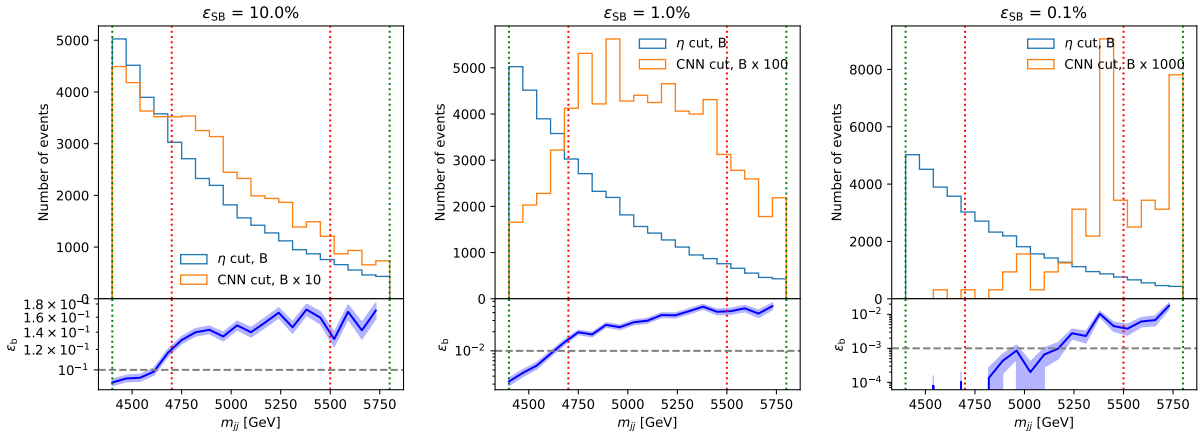
5. Evaluate the difference between the number of events to estimate the number of signal events and compute the sensitivity after the CWoLa cut.

We assume the real background is similar to the simulated background and the network performance of the real and simulated events is similar.

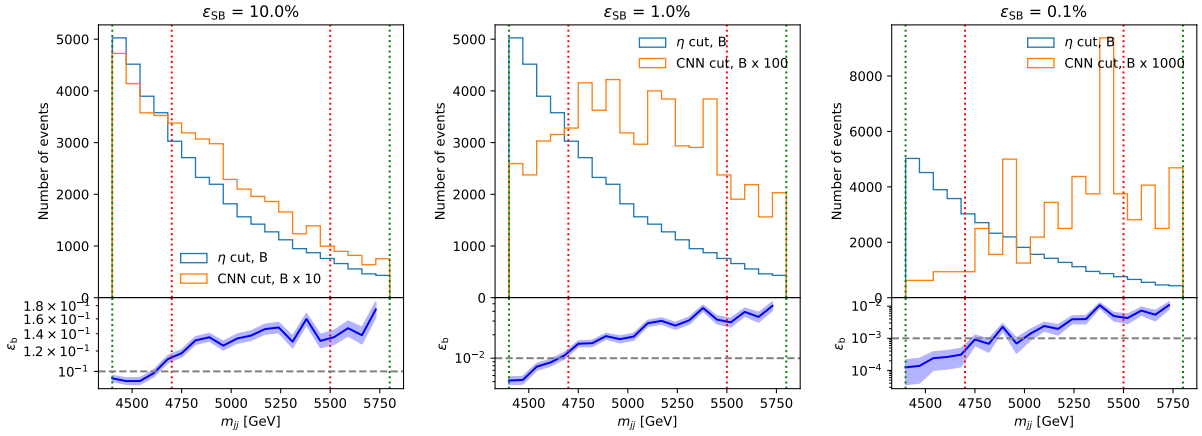
Figure 127 shows the sensitivity improvement evaluated by different methods. The "true label" is our previous method, where we use another testing dataset in the signal region to evaluate the sensitivity improvement. The background subtraction and the previous method



(a) Original, $S/\sqrt{B} = 0$

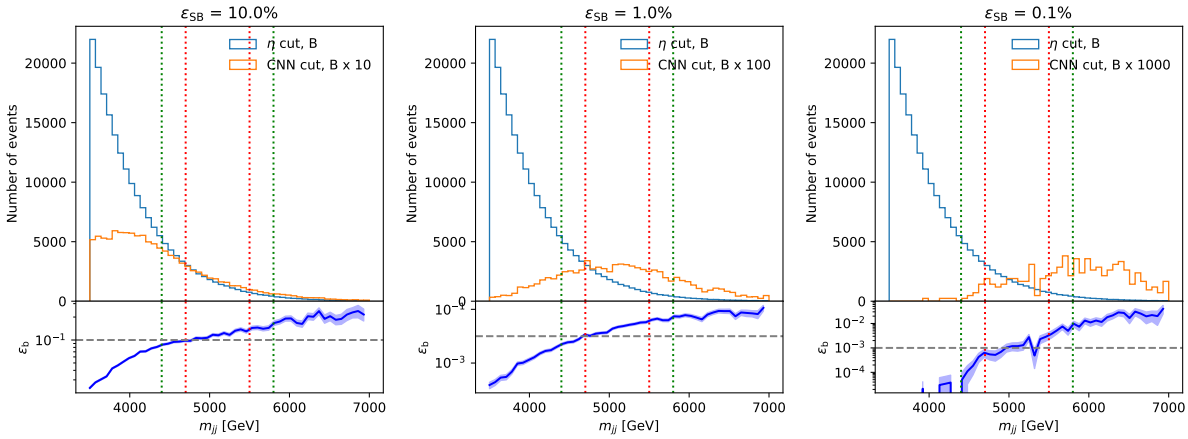


(b) +3 jet rotation, $S/\sqrt{B} = 0$

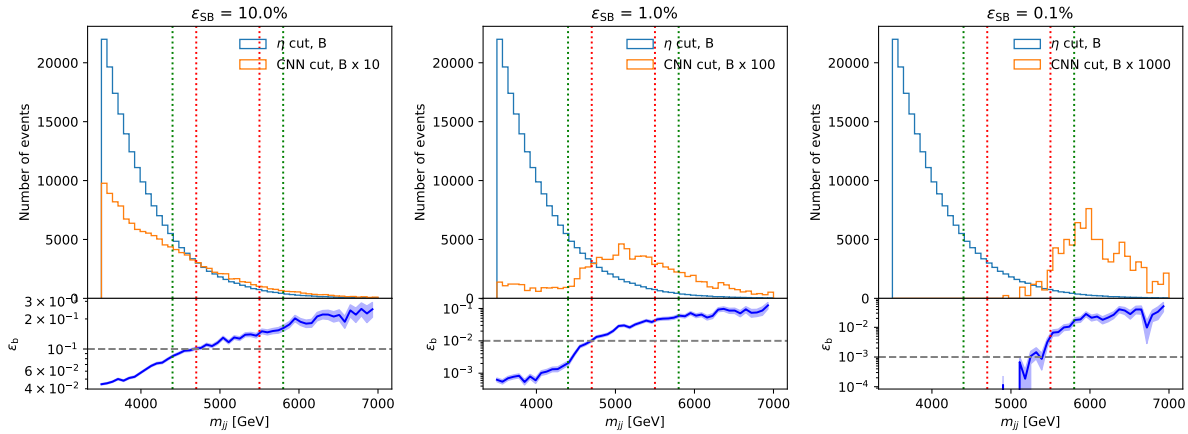


(c) +3 p_T smearing + jet rotation, $S/\sqrt{B} = 0$

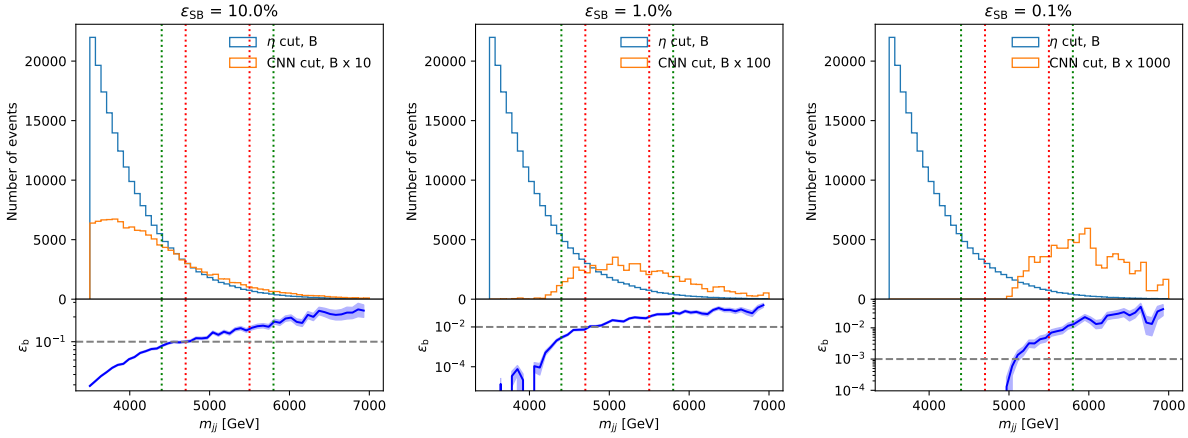
Figure 125: The m_{jj} distribution before and after the CWoLa CNN selection. The signal region is between the red dotted lines. The sideband region is between the green dotted lines and excludes the signal region. "B" stands for the background samples.



(a) Original, $S/\sqrt{B} = 0$



(b) +3 jet rotation, $S/\sqrt{B} = 0$



(c) +3 p_{T} smearing + jet rotation, $S/\sqrt{B} = 0$

Figure 126: The m_{jj} distribution before and after the CWoLa CNN selection. The signal region is between the red dotted lines. The sideband region is between the green dotted lines and excludes the signal region. "B" stands for the background samples.

give similar results for $\varepsilon_b = 1.0\%, 0.1\%$. There is much difference for $\varepsilon_b = 10\%$. Even if we consider the no-signal case the sensitivity is close to 5 after the CWoLa cut. Figure 128 shows the sensitivity improvement evaluated by the background subtraction method. The augmented datasets perform better with lower background efficiencies.

Figure 129 shows the background efficiencies of training datasets. The passing rate of training sets is consistently higher than testing sets. This satisfied our expectations, the neural network would prefer to choose the training samples.

4.51 Mitigate the over-fitting on training dataset

Figure 129 shows the over-fitting issue on the training samples. This would lead to a fake sensitivity. We test the Dropout and L2 regularization techniques to mitigate the over-fitting issue.

First, the Dropout technique is only applied on DNN layers, but there is no difference in the training results. Thus, the Dropout is always applied on all layers with the same dropout rate. Figure 130 shows the training results of different dropout rates. For dropout rates 0.25 and 0.35, they can mitigate the over-fitting issue and perform similarly. The training would fail for the dropout rate of 0.50.

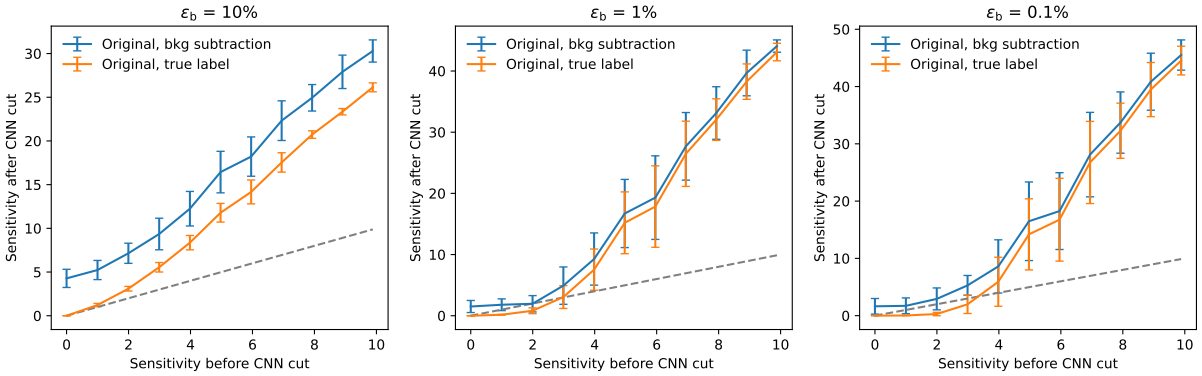
In the following training, we set the dropout rate to 0.35. Figure 131 shows the training results with L2 regularization. Here, the L2 regularization is only applied on CNN layers. When we used it on all layers the training would fail. It seems that L2 regularization can not further reduce the over-fitting issue.

Figure 132 shows the sensitivity improvement with the Dropout technique. We use the background subtraction to evaluate the sensitivity. Figure 133 shows the sensitivity improvement with different training dataset sets. The augmented datasets perform better and have smaller fluctuations.

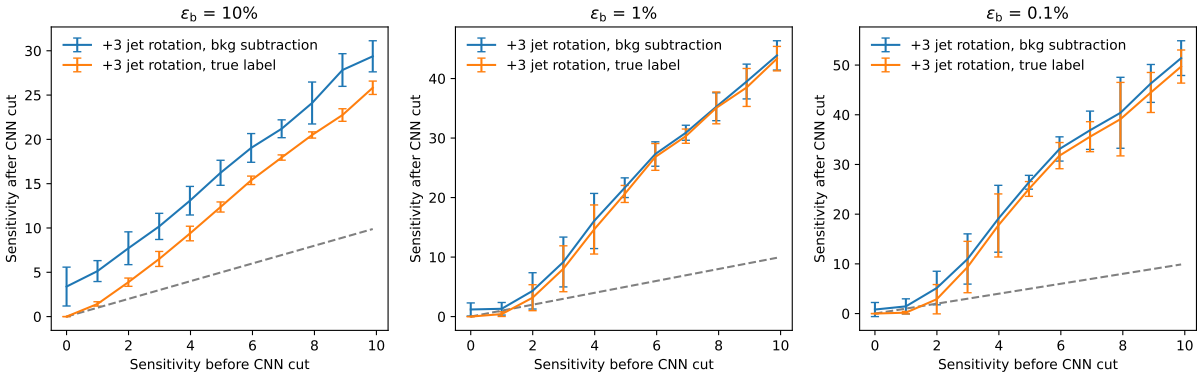
Figure 134 shows the background efficiencies of training datasets. The passing rate of training sets is consistently higher than that of testing sets but lower than Figure 129.

4.52 K-fold Cross-Validation

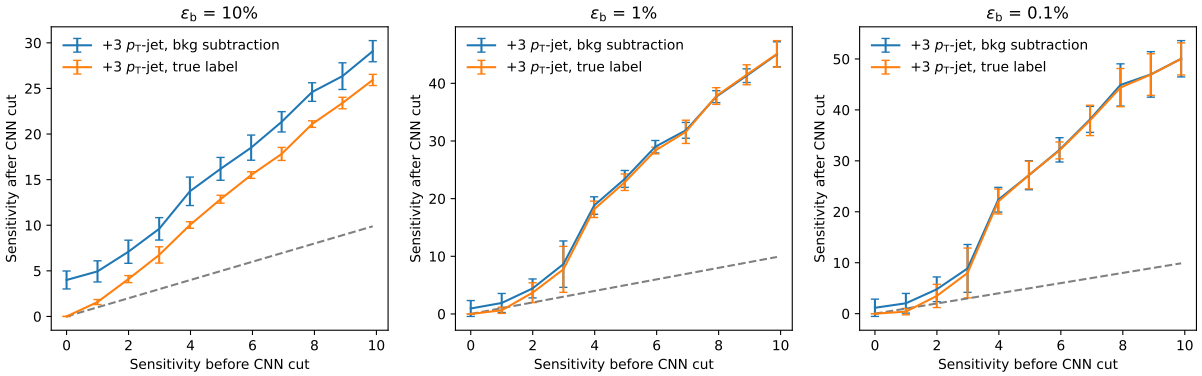
K-fold cross-validation is a technique used to evaluate the performance of a machine-learning model. The k-fold method divides the dataset into k equally sized subsets, or "folds." For each fold, treat one fold as the test set and the remaining $k-1$ folds as the training set. After completing the training for all k folds, calculate the average performance metric. This average provides a more robust estimate of the model's performance than a single train-test split.



(a) Original



(b) +3 jet rotation



(c) +3 p_{T} smearing + jet rotation

Figure 127: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

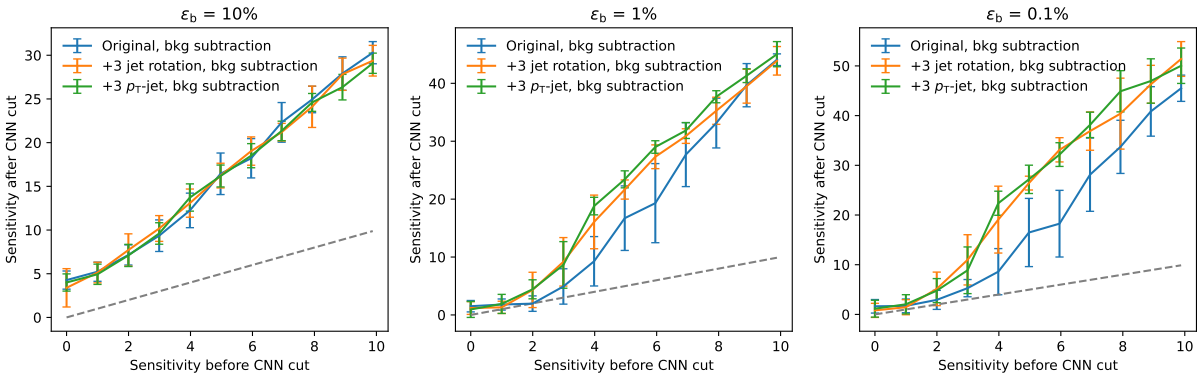


Figure 128: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

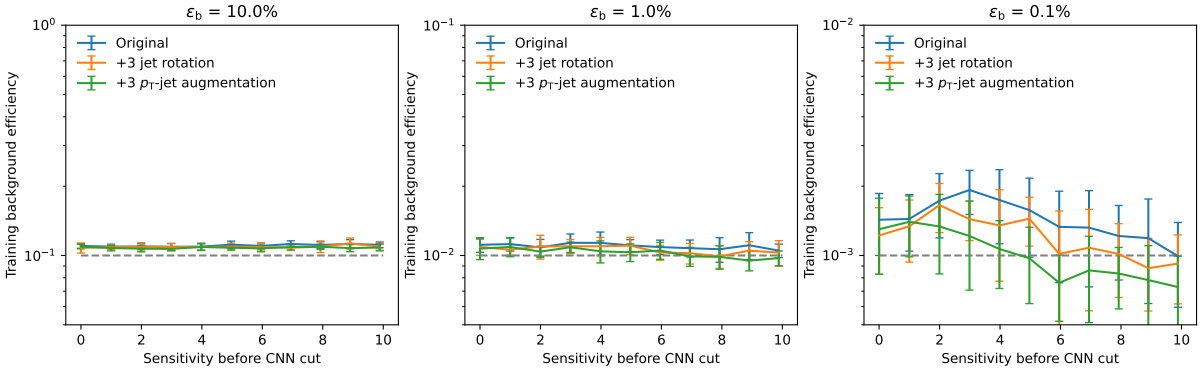
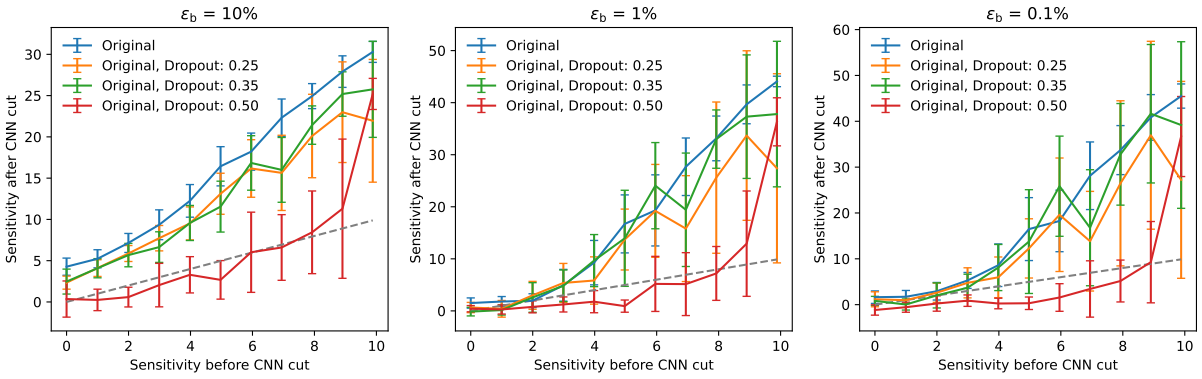
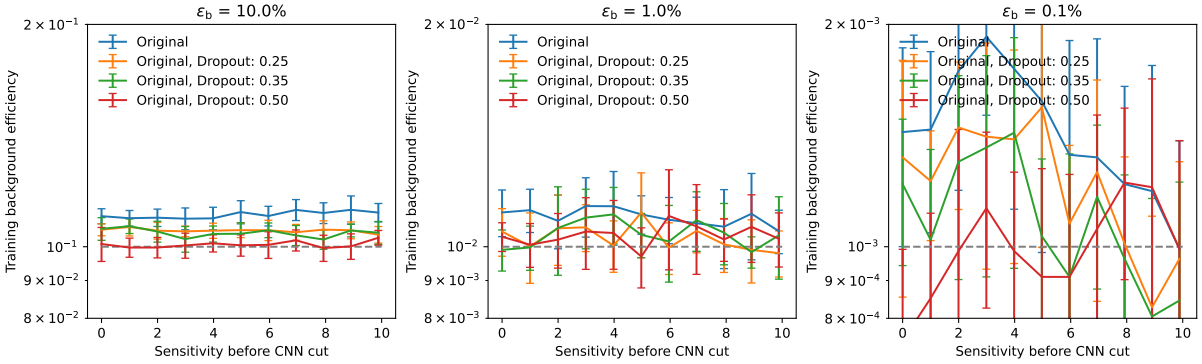


Figure 129: The training background efficiency of CWoLa CNN selection. The dashed grey line is 1 representing the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

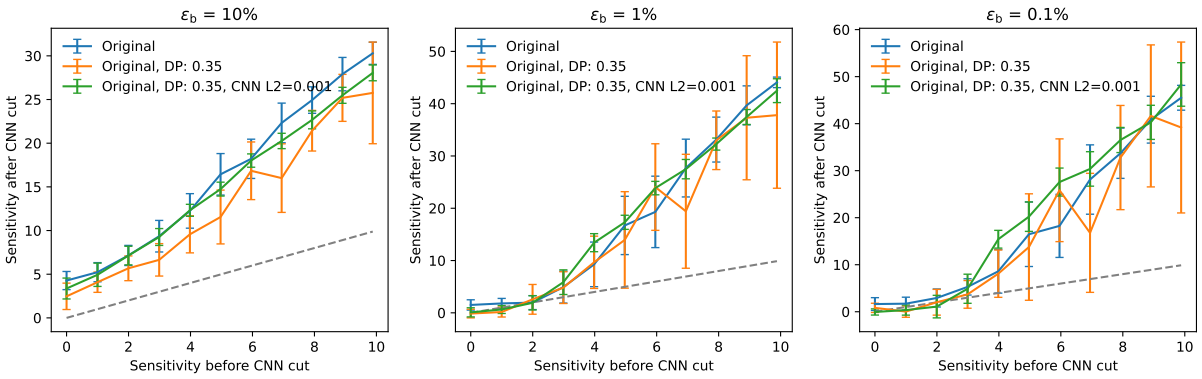


(a) Original

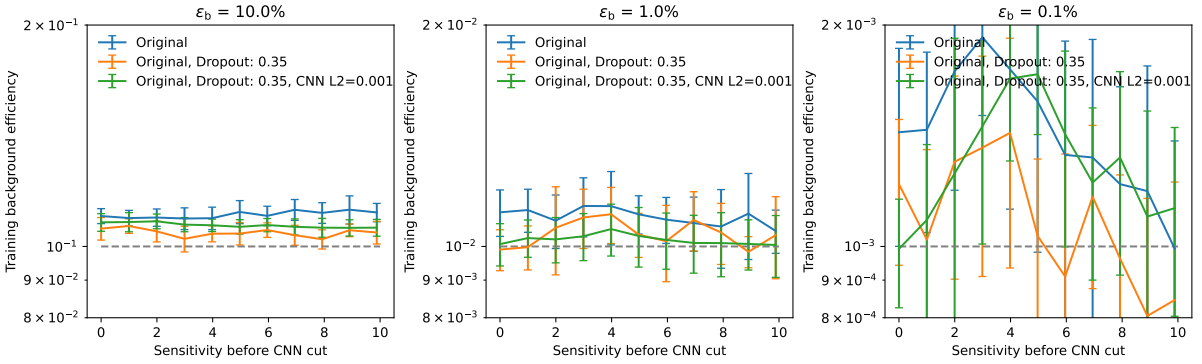


(b) +3 jet rotation

Figure 130: (a) The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. (b) The training background efficiencies of CWoLa CNN selection. The dashed grey line represents the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

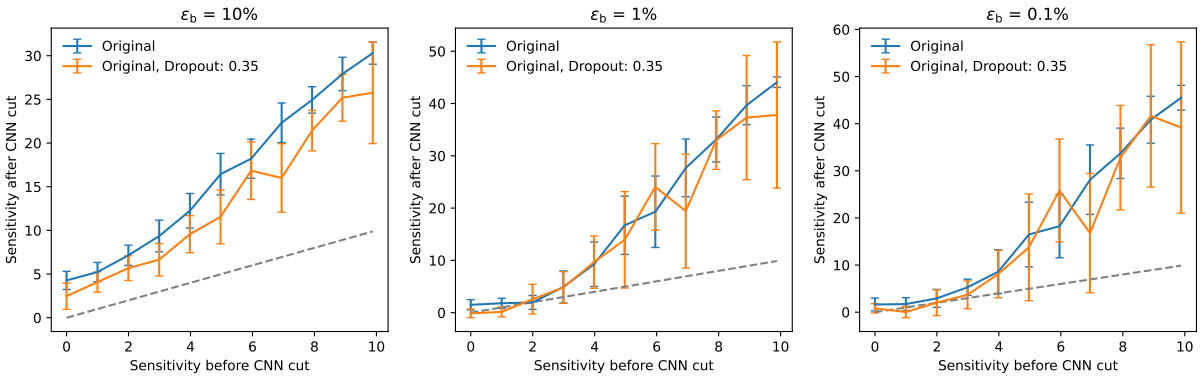


(a) Sensitivity improvement

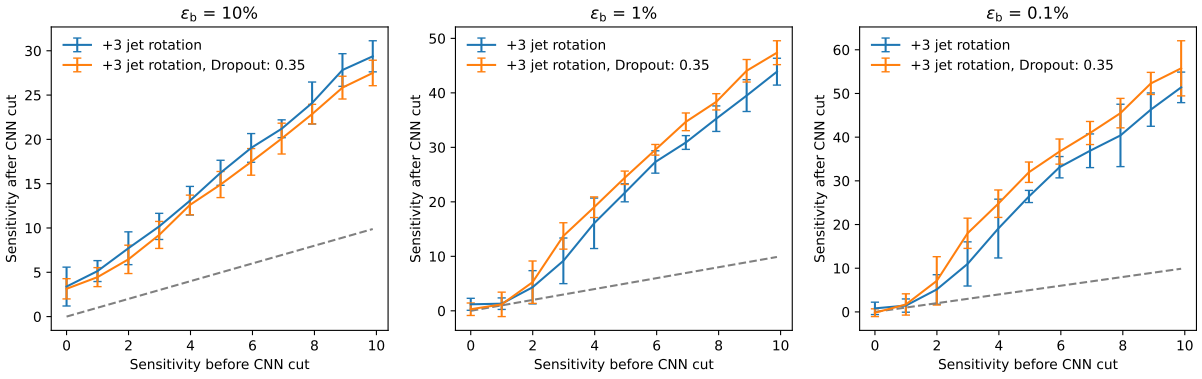


(b) Training background efficiency

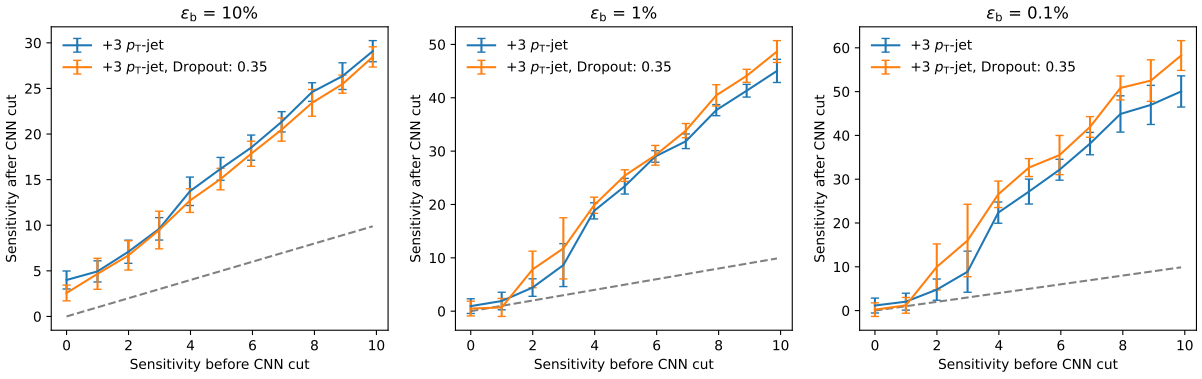
Figure 131: (a) The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. (b) The training background efficiencies of CWoLa CNN selection. The dashed grey line represents the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.



(a) Original



(b) +3 jet rotation



(c) +3 p_{T} smearing + jet rotation

Figure 132: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

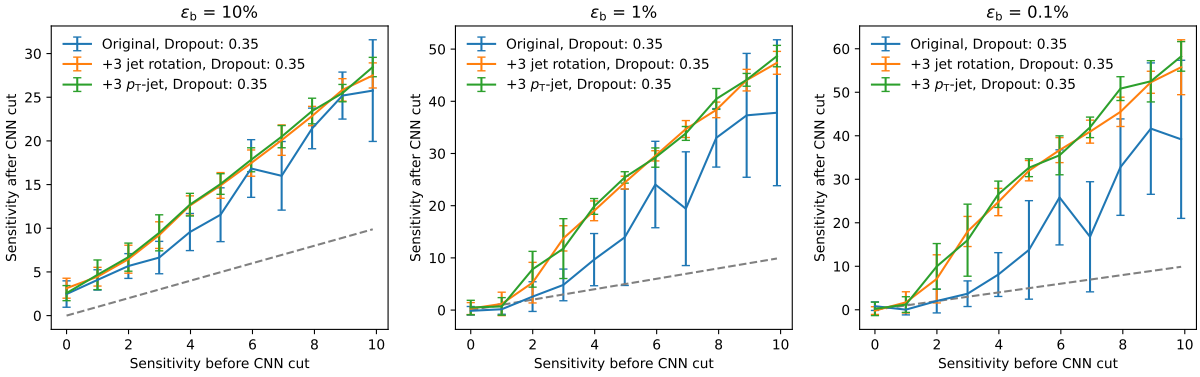


Figure 133: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

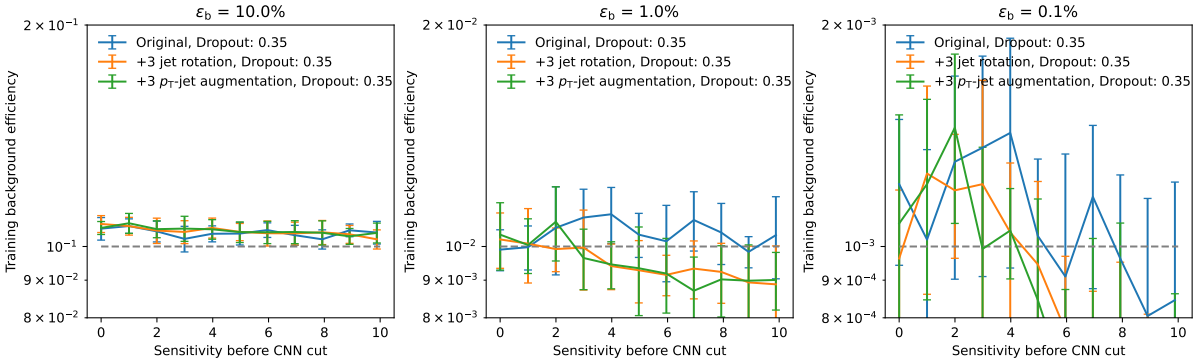
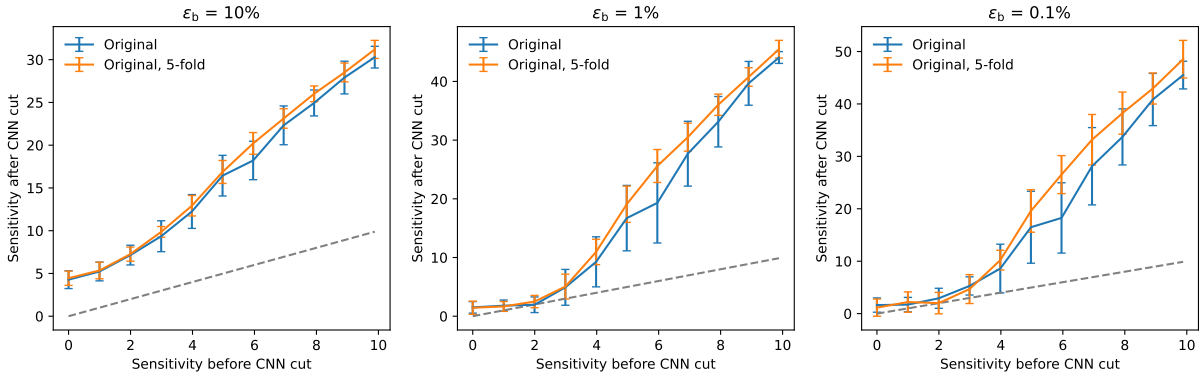


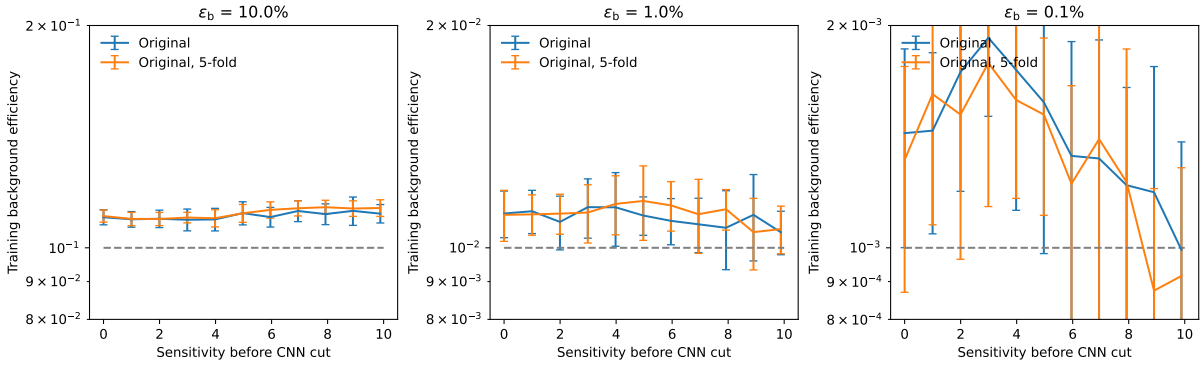
Figure 134: The training background efficiency of CWoLa CNN selection. The dashed grey line is 1 representing the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

We set $k = 5$ for our CWoLa training. After completing the training for all k folds, we evaluate the average event scores as the final output.

Figure 135 shows the training results with the k-fold technique. The training performance with and without the k-fold technique are similar. The k-fold technique only reduces the training fluctuation.



(a) Sensitivity improvement



(b) Training background efficiency

Figure 135: (a) The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. (b) The training background efficiencies of CWoLa CNN selection. The dashed grey line represents the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

4.53 L1 regularization

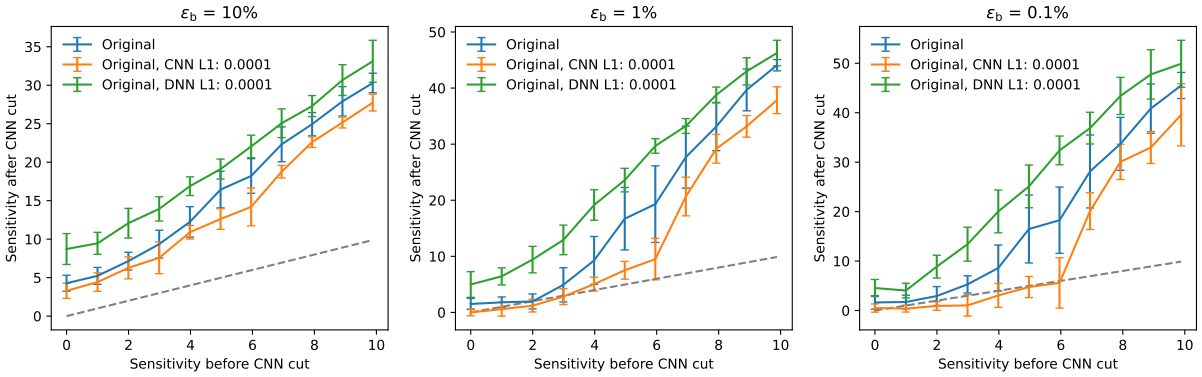
L1 regularization is a technique to prevent overfitting by adding a penalty term to the loss function. L1 regularization adds the absolute values of the coefficients (weights) to the

loss function:

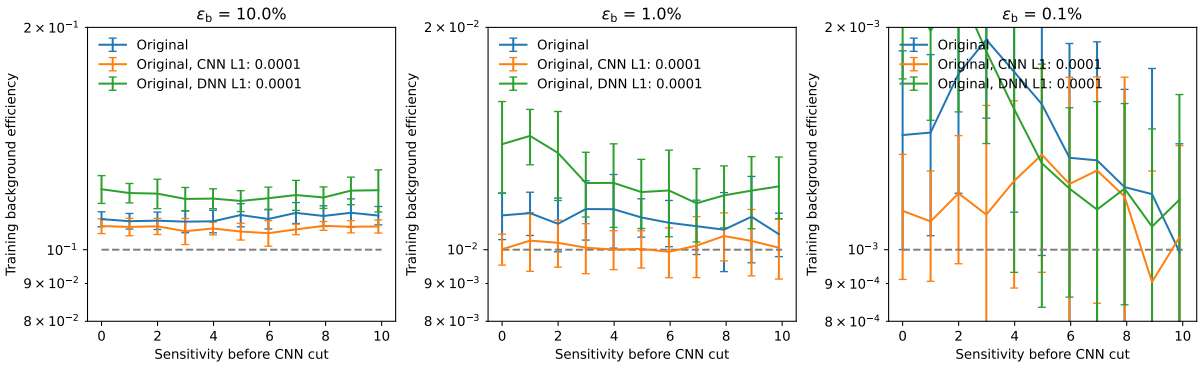
$$\text{Loss} = \text{Original Loss} + \lambda \sum_i |w_i| \tag{9}$$

where w_i are the model's coefficients and λ is a hyperparameter that controls the strength of the penalty term.

Figure 136 shows the training results with the L1 regularization technique. The "CNN L1: 0.0001" means the L1 regularization is applied on all parameters in the CNN layers with $\lambda = 0.0001$. The "CNN L1: 0.0001" can mitigate the overfitting, while the "DNN L1: 0.0001" worsens the situation.



(a) Sensitivity improvement



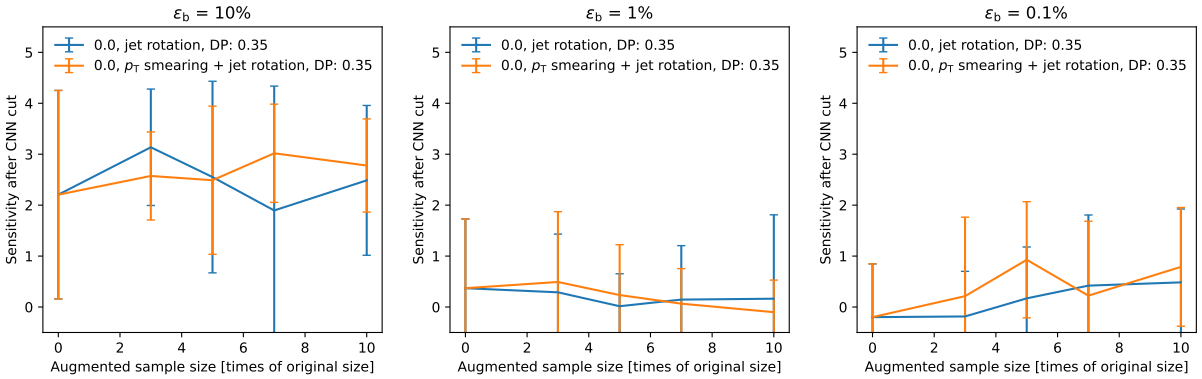
(b) Training background efficiency

Figure 136: (a) The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. (b) The training background efficiencies of CWoLa CNN selection. The dashed grey line represents the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

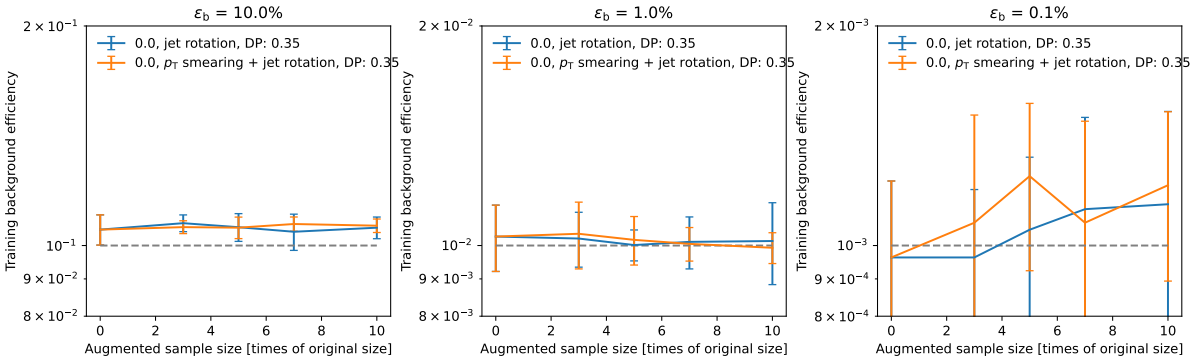
4.54 Asymptotic behaviour of no-signal neural network

We enlarged the training sample size to identify whether the more augmented samples would improve the situation. In Section 4.36, we found the sensitivity improvement increases with the training sample size. However, we didn't perform the same exercise on no-signal case. We train the CNN on larger datasets without signal events.

Figure 137 shows the training results with different sample sizes on the no-signal network. The Dropout technique is applied on all layers with a dropout rate 0.35. It seems that the training results are independent of the augmented sample size.



(a) Sensitivity improvement



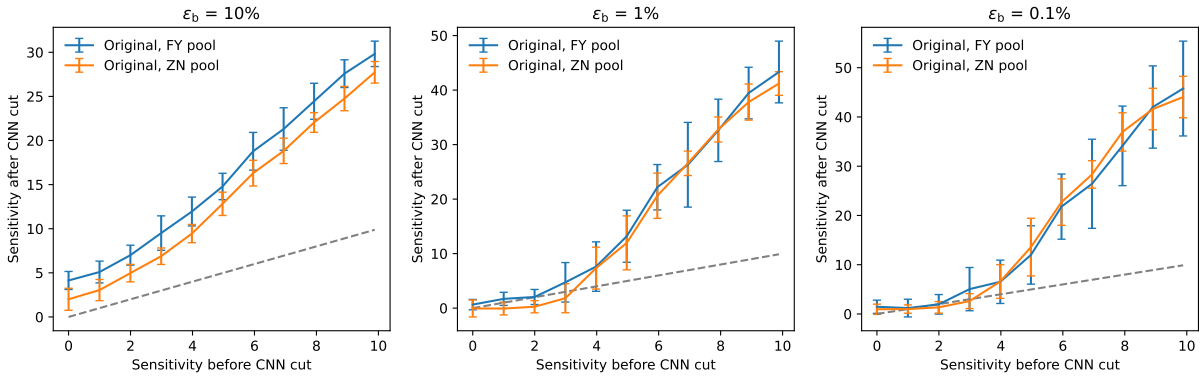
(b) Training background efficiency

Figure 137: (a) The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. (b) The training background efficiencies of CWoLa CNN selection. The dashed grey line represents the testing background efficiency. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

4.55 Change the training dataset

I used ZN's data pool for training to make sure I could reproduce similar results with these datasets. First, I split ZN's pool into training and validation sets, then resample from these two datasets in each training. The testing samples are the same as my previous tests.

Figure 138 shows the training results with ZN's data pool. The mean value of the ZN pool at sensitivity 0 is 2.00 which is lower than the 4.27 of the FY pool.



(a) Sensitivity improvement

Figure 138: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

To examine whether I accidentally choose very bad training data from the FY pool, I perform 100 training with 100 different random seeds.

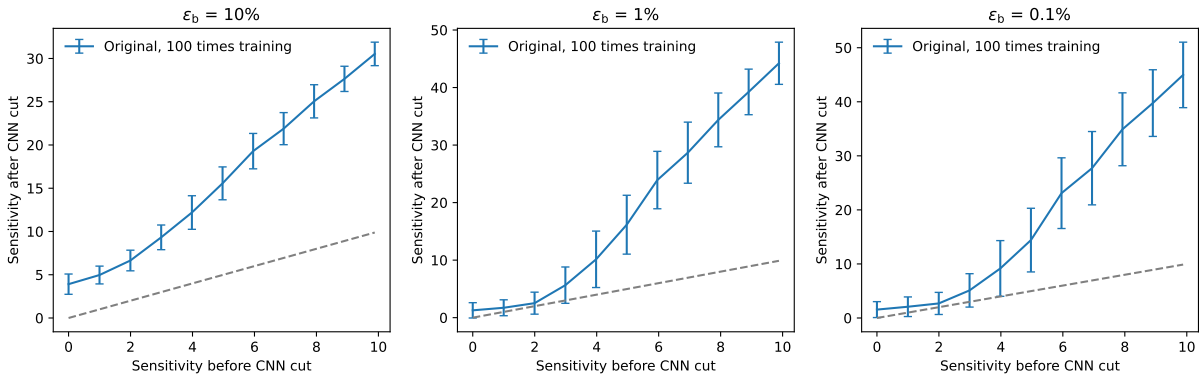
Figure 139 shows the training results with 100 times training. The mean value of 100 times training at sensitivity 0 is 3.91 which is close to the previous results.

There could be something wrong with my testing dataset. Therefore, I would prepare another testing dataset and perform the same training with the new testing samples.

4.56 Another testing dataset

The same testing dataset is used in previous exercises. To make sure whether I accidentally chose a bad testing dataset, I prepared other testing sets and performed the same training with the new testing samples.

There are three testing datasets. Testing set 1 is the old dataset with 10k signals and 10k background events. Testing set 2 is a new dataset with 10k signals and 10k background events. Testing set 3 is another new dataset with 20k signals and 20k background events.



(a) Sensitivity improvement

Figure 139: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 100 times training. We use different datasets for each training.

Figure 140 shows the training results with different testing datasets. We focus on the results at sensitivity 0. The sensitivity improvement is 4.23 ± 1.04 for testing set 1, 2.78 ± 1.06 for testing set 2, 2.40 ± 0.81 for testing set 3. It seems that the sensitivity improvement would depend on the testing datasets.

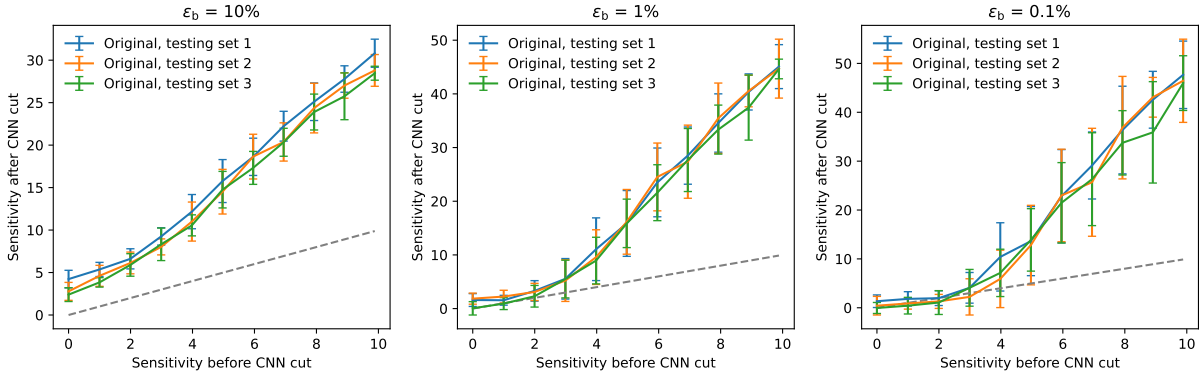


Figure 140: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

Figure 141 shows the training results with 100 times training and the new testing datasets. The sensitivity improvement is 3.82 ± 1.23 for testing set 1, and 2.98 ± 1.20 for testing set 2. This shows that different training samples would impact the sensitivity

improvement, but roughly speaking the same testing dataset would give similar sensitivity improvement. The most reasonable approach is preparing a larger testing set to reduce the testing fluctuation from different samples.

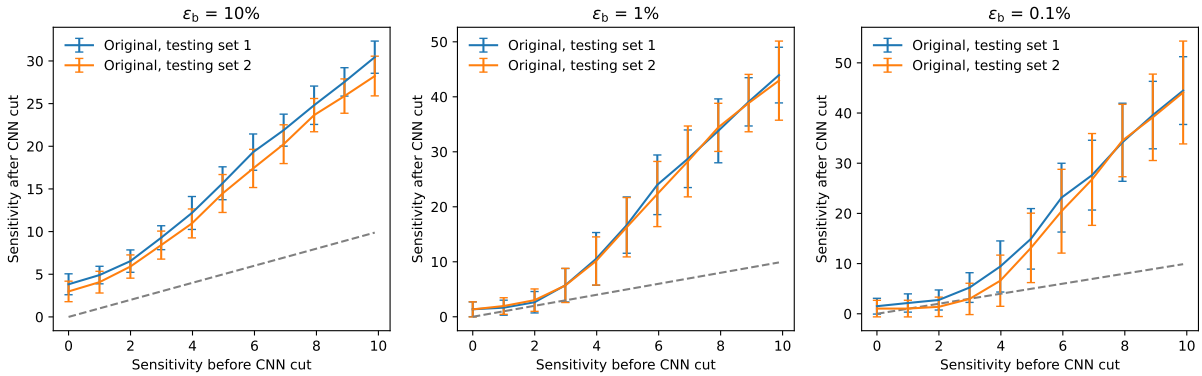


Figure 141: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 100 times training. We use different datasets for each training.

4.57 Enlarge the testing dataset

In Section 4.56, we noticed that the sensitivity improvement depends on the testing sets. We prepare a larger testing set to reduce the testing fluctuation from different samples. This new testing set contains 100k signals and 100k background events.

Figure 142 shows the training results with larger testing datasets. We focus on the results at sensitivity 0 with $\varepsilon_b = 10\%$. The sensitivity improvement is 4.23 ± 1.04 for testing set 1, 2.63 ± 0.92 for 100k + 100k testing set. This value is consistent with testing sets 2 and 3 shown in Figure 140.

Figure 143 shows the training results with ZN's data pool. The mean value of the ZN pool at sensitivity 0 with $\varepsilon_b = 10\%$ is 0.98 ± 2.36 which is lower than the 2.63 ± 0.92 of the FY pool.

Figure 147 shows the sensitivity improvement with different training dataset sets. The augmented datasets perform better. At sensitivity 0, the augmented sets perform similarly to the original datasets.

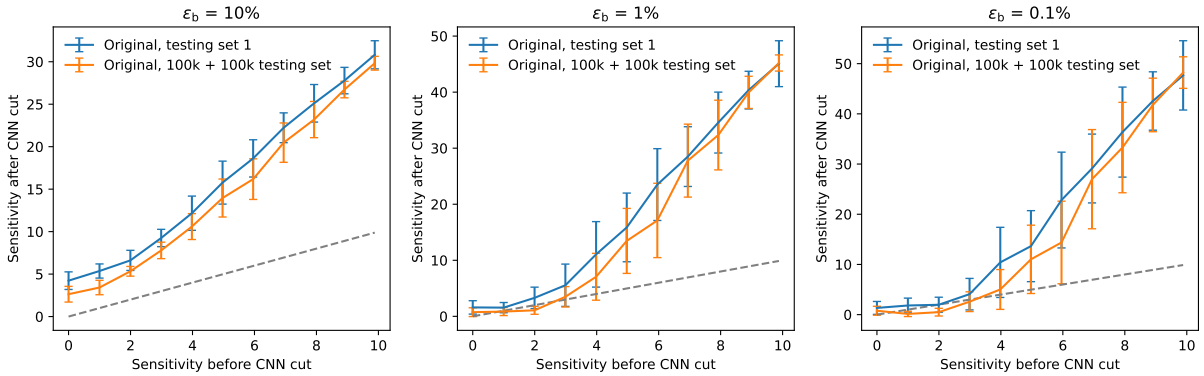
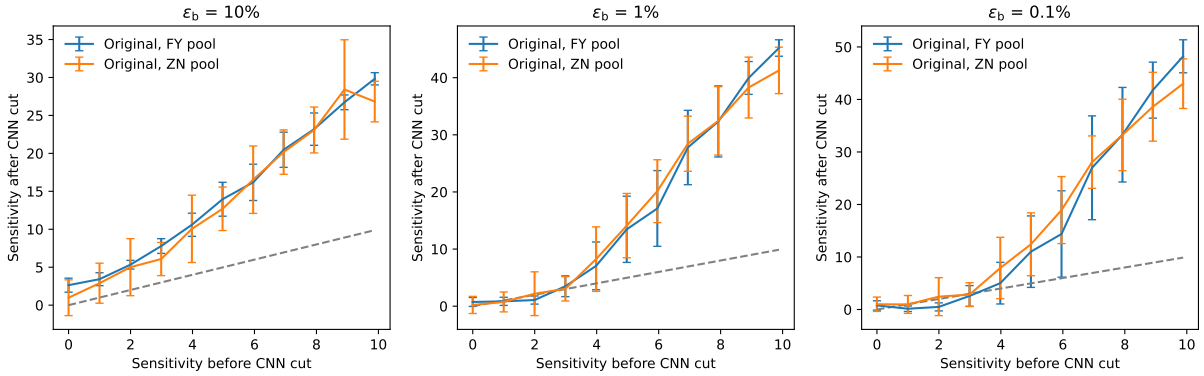


Figure 142: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.



(a) Sensitivity improvement

Figure 143: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

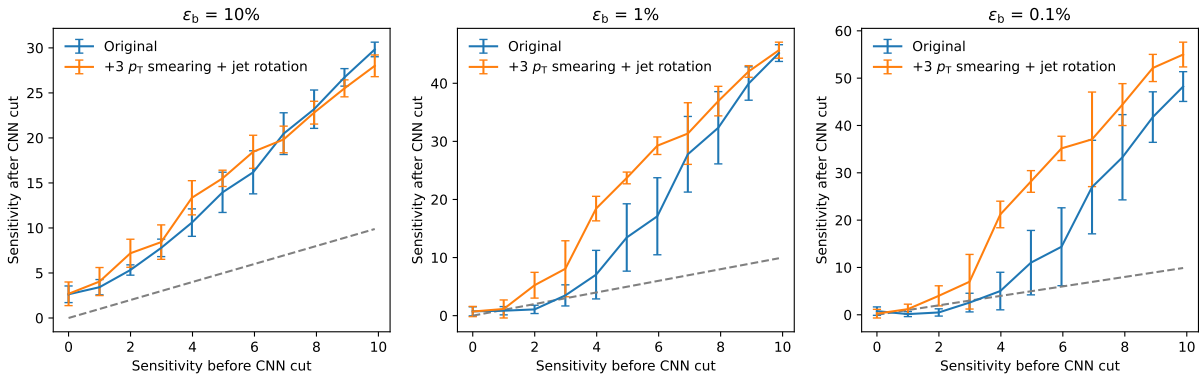


Figure 144: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

4.58 Jet image normalization

To mitigate the sculpting effect, we normalize the jet image by the invariant mass m_{jj}. This technique could reduce the difference between the SR and SB regions.

First, compute the total invariant mass m_{jj} for the event passing the pre-selection cuts. Then, use the m_{jj} to normalize the transverse momentum p_{T} of each jet constituent:

$$p'_{\mathrm{T}} = p_{\mathrm{T}} \times \frac{5100 \text{ GeV}}{m_{jj}} \tag{10}$$

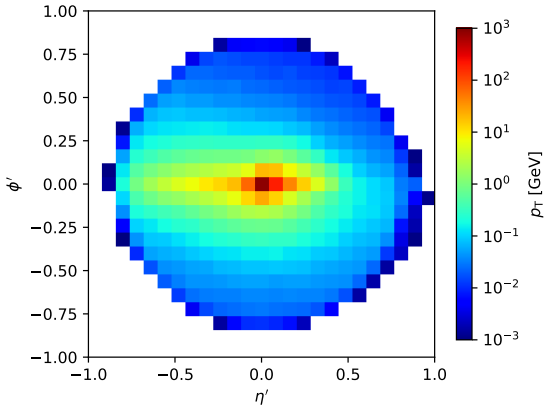
where p'_{T} is the normalized transverse momentum of the jet constituent, and the value 5100 GeV is the center of the signal region. The following steps are the same as the previous procedure, i.e., perform the preprocessing and apply the data augmentation.

Figure 145 is the original and p_{T} normalized jet images. Here, we show the jet images of background events in the SB region. The p_{T} normalized and original cases have a similar distribution, but the values are not the same.

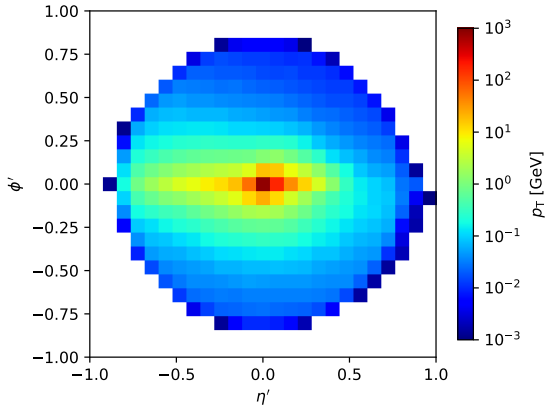
Figure 146 shows m_{jj} distributions and background passing rates for the training dataset. The number of events is normalized by the cross-section. The ε_{b} is the number of events after the CNN cut divided by the value before the CNN cut. The grey dashed line equals the background passing rate at the sideband region. The band is the standard deviation of the ε_{b}, evaluated from the Equation 7.

There are obvious sculpting issues for all the background efficiencies. The p_{T} normalization makes the situation worse than previous cases (Figure 125).

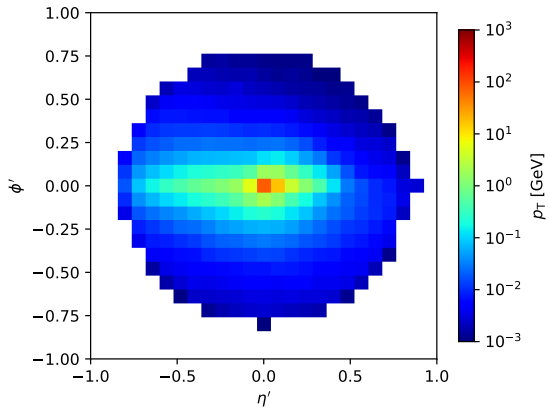
Figure 147 shows the sensitivity improvement with p_{T} normalization technique. The training thresholds are much higher than the original cases. However, when we focus on the



(a) p_T normalized

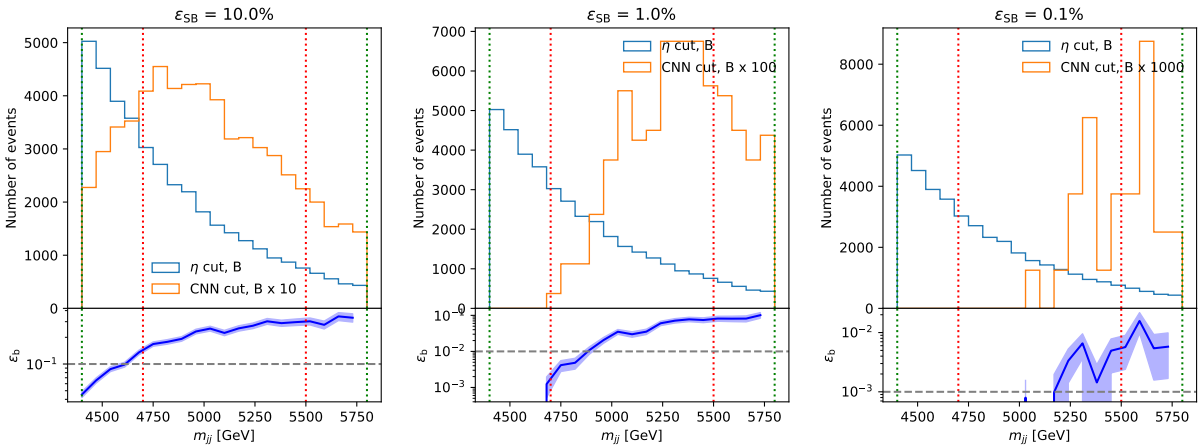


(b) Original



(c) Difference

Figure 145: The p_T normalized, original, and their difference average jet images.



(a) Original, p_{T} normalized, $S/\sqrt{B} = 0$

Figure 146: The m_{jj} distribution before and after the CWoLa CNN selection. The signal region is between the red dotted lines. The sideband region is between the green dotted lines and excludes the signal region. "B" stands for the background samples.

results at sensitivity 0 with $\varepsilon_{\mathrm{b}} = 10\%$. The sensitivity improvement is 2.63 ± 0.92 for original method, 0.71 ± 1.00 for p_{T} normalization.

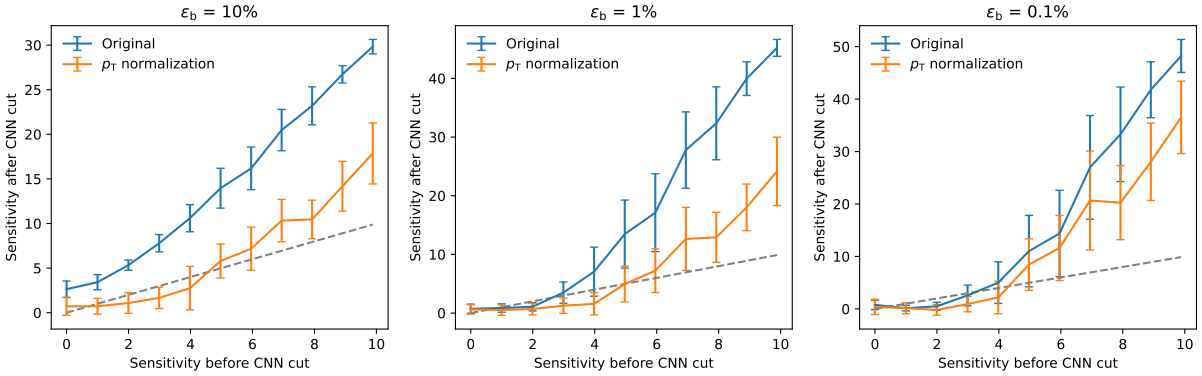


Figure 147: The sensitivities before and after the CWoLa CNN selection. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. We use different datasets for each training.

References

[1] J. H. Collins, K. Howe, and B. Nachman, "Anomaly Detection for Resonant New Physics with Machine Learning," *Phys. Rev. Lett.*, vol. 121, no. 24, p. 241803, 2018.

[2] B. M. Dillon, L. Favaro, F. Feiden, T. Modak, and T. Plehn, "Anomalies, Representations, and Self-Supervision," 1 2023.

[3] H. Beauchesne, Z.-E. Chen, and C.-W. Chiang, "Improving the performance of weak supervision searches using transfer and meta-learning," *JHEP*, vol. 02, p. 138, 2024.