Note

Feng-Yang Hsieh

1 Signal

We consider a simple extension of the standard model (SM) [1], which includes a vector-like dark fermion $(\overline{\chi}, \chi)$ and a complex singlet scalar S. A signature of CP violation could come from the Higgs-to-Higgs decays, $h_3 \to h_2 h_1$, where $h_3/h_2/h_1$ are the heaviest scalar, the second heaviest scalar, and the SM-like 125 GeV Higgs, respectively.

The signal process is the triple production of 125 GeV Higgs bosons via the gluon fusion:

$$gg \to h_3 \to h_2 h_1 \to h_1 h_1 h_1$$

The Higgs boson h_1 would further decay to the $b\bar{b}$ pair. We consider the banchmark point 1 (BP1), where $m_{h_3} = 420 \text{ GeV}, m_{h_2} = 280 \text{ GeV}, m_{h_1} = 125 \text{ GeV}$. This process is generated at $\sqrt{s} = 13$ TeV. Following are the MadGraph scripts for generating signal samples:

```
import model cxSM_VLF_EFT
generate g g > h h h
output MG5/gghhh_bsm
launch MG5/gghhh_bsm

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

set param_card mh1 125
set param_card mh2 280
set param_card mh3 420
set param_card theta12 0.73
set param_card theta13 1.67079632679
set param_card theta23 -0.73
set param_card vs 200
set param_card delta2 0
set param_card Rdelta3 0
set param_card Idelta3 -3.5
set param_card b2 0
set param_card Rc1 0
set param_card Ic1 0
set param_card Rc2 0
set param_card Ic2 0
set param_card Rd3 0
set param_card Id3 0
set param_card msq -5033.406281907266
set param_card lam 0.13850082540690806
set param_card Rdelta1 -47.561525227572744
set param_card Idelta1 853.05384671134
set param_card Rb1 -70476.6380004269
set param_card Ib1 -30486.140015405872
set param_card Rd1 -2.562109886826132
set param_card Id1 2.257859679994403
set param_card d2 6.340799300844676
set param_card gh1ggr -0.00005478952893059635
set param_card gh1gagar -0.00003270447254456052
set param_card gh1Zgar -0.00005871986046374793
set param_card gh2ggr -1.4279972541632635e-7
set param_card gh2gagar -8.237715486808595e-8
set param_card gh2Zgar -1.3984990232267825e-7
set param_card gh3ggr -6.031835872118092e-6
set param_card gh3gagar -1.1377279177203616e-6
set param_card gh3Zgar -2.2999597941282603e-6

set param_card decay 102 auto
set param_card decay 103 auto

set run_card nevents 100000
set run_card ebeam1 6500.0
set run_card ebeam2 6500.0

set run_card ptb 24
set run_card etab 2.6

set spinmode none
decay h > b b~

done
```

2 SPANet pairing

We employ the novel neural network structure Spa-Net [2, 3, 4] to identify the correct pairings among the jets in the final states.

2.1 Training dataset preparation

Preselection: ≥ 6 jets with transverse momentum $p_{\mathrm{T}} \geq 25$ GeV in range $|\eta| < 2.5$.

The input features for the Spa-Net are a list of jets, each represented by its 4-component vector $(p_{\mathrm{T}}, \eta, \phi, m)$ as well as a boolean b-tag. We only keep each event's 15 highest p_{T} jets. We define the correct jet assignments for each event by matching the jets to the simulated truth quarks within an angular distance of $\Delta R < 0.4$. Such an event will be dropped if a simulated truth quark is matched to more than one jet. Furthermore, some simulated truth quarks may not be matched to any jet, so the event will not be used in training either.

After the selection and matching, we could obtain the following results from 1M events:

- Total sample size: 522,899
- 1h sample size: 184,769
- 2h sample size: 161,476
- 3h sample size: 94,464

Here, the 1h sample is where we could define the correct jet assignments for 1 Higgs boson.

2.2 Training results

- Training sample:
    - Total sample size: 470,609
    - 1h sample size: 166,490
    - 2h sample size: 145,309
    - 3h sample size: 84,913
    - 5% used on validation
- Testing sample:
    - Total sample size: 52,290
    - 1h sample size: 18,279
    - 2h sample size: 16,167
    - 3h sample size: 9,551

Some useful definitions for evaluating jet assignment performance:

- Event Efficiency

$$\epsilon^{\text{event}} \equiv \frac{\text{number of events with and all Higgs are correctly identified}}{\text{number of events}} \tag{1}$$

- Higgs Efficiency

$$\epsilon^{\text{h}} \equiv \frac{\text{number of correctly identified Higgs}}{\text{number of identifiable Higgs}} \tag{2}$$

The training results are shown in Table 1.

Table 1: SPA-NET pairing efficiencies on 3h events.

N_{Jet}	Event Fraction	Event Efficiency	Higgs Efficiency
$= 6$	0.077	0.532	0.650
$= 7$	0.057	0.345	0.536
≥ 8	0.052	0.237	0.452
Total	0.186	0.375	0.548

3 χ^2 pairing

χ^2 method considers all possible combinations of final jets and selects the configuration that minimizes the mass difference between Higgs candidates and SM Higgs, i.e., minimizes this:

$$\chi^2 = [m(j_1 j_2) - m_h]^2 + [m(j_3 j_4) - m_h]^2 + [m(j_5 j_6) - m_h]^2 \tag{3}$$

where $m(j_i j_j)$ is the invariant mass of jet i, j and $m_h = 125$ GeV.

Table 2 is the performance of the χ^2 method.

Table 2: χ^2 pairing efficiencies on 3h events.

N_{Jet}	Event Fraction	Event Efficiency	Higgs Efficiency
$= 6$	0.077	0.403	0.450
$= 7$	0.057	0.158	0.281
≥ 8	0.052	0.000	0.077
Total	0.186	0.215	0.294

4 Estimate cross-section of background process

Besides the 6 b background, we need to consider the backgrounds that come from the mis-tagging of light jets or charm-jets to b-jets. We assume that the probability of a charm-jet being misidentified as b-jet is $\mathcal{P}_{c\to b} = 0.1$ and that of light jets is $\mathcal{P}_{j\to b} = 0.01$. The b-tagging efficiency is assumed to be $\mathcal{P}_{b\to b} = 0.7$.

Table 3 shows the cross-section computed from MadGraph and the cross-section times the mis-tagging probabilities $\mathcal{P}_{c\to b}$ and $\mathcal{P}_{j\to b}$. Table 4 shows the same results with kinetic cuts. We require the transverse momentum p_{T} of each jet greater than 24 GeV and in the range $|\eta| < 2.6$ at the MadGraph level. The $6b$ process contributes much more than the processes containing charm jets and light jets.

5 Compute pairing efficiency

To understand the pairing performance with different pairing methods, we compute how many events where 1h/2h/3h bosons are reconstructed correctly.

The pairing performance of Spa-Net are shown in Table 5. Table 6 is the performance of the χ^2 method. For both cases, we found the number of events where only two Higgs are

Table 3: The cross-sections of $6b$ and mis-tagging background processes. The cross-sections are computed from the MadGraph at $\sqrt{s} = 13$ TeV.

process	σ (pb)	$\sigma \times \mathcal{P}$(tagging efficiency) (pb)
$(b\bar{b})(b\bar{b})(b\bar{b})$	2.53×10^3	2.97×10^2
$(b\bar{b})(b\bar{b})(c\bar{c})$	2.72×10^2	6.54×10^{-1}
$(b\bar{b})(c\bar{c})(c\bar{c})$	3.73×10^1	1.83×10^{-3}
$(b\bar{b})(b\bar{b})(jj)$	7.44×10^4	1.79

Table 4: The cross-sections of $6b$ and mis-tagging background processes. The cross-sections are computed from the MadGraph at $\sqrt{s} = 13$ TeV. We require the transverse momentum p_{T} of each jets greater than 24 GeV in range $|\eta| < 2.6$.

process	σ (fb)	$\sigma \times \mathcal{P}$(tagging efficiency) (fb)
$(b\bar{b})(b\bar{b})(b\bar{b})$	9.63×10^2	113.35
$(b\bar{b})(b\bar{b})(c\bar{c})$	1.67×10^3	4.02
$(b\bar{b})(c\bar{c})(c\bar{c})$	1.06×10^3	5.19×10^{-2}
$(b\bar{b})(b\bar{b})(jj)$	4.16×10^5	9.98
$(b\bar{b})(jj)(jj)$	1.50×10^7	7.73×10^{-2}

paired correctly is tiny, which means if we can pair two Higgs bosons correctly, then we have a high chance to pair the final Higgs correctly.

Note that Higgs Efficiencies of Spa-Net are inconsistent with Table 1. This issue needs more checking.

Table 5: Spa-Net pairing efficiencies on different categories.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.532	0.000	0.119	0.348	0.572
$= 7$	0.345	0.021	0.166	0.469	0.414
≥ 8	0.237	0.022	0.186	0.554	0.314
Total	0.375	0.014	0.156	0.455	0.436

Table 6: χ^2 pairing efficiencies on different categories.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.403	0.000	0.143	0.455	0.450
$= 7$	0.158	0.070	0.228	0.544	0.281
≥ 8	0.000	0.000	0.231	0.769	0.077
Total	0.215	0.022	0.194	0.570	0.294

6 SPANet pairing and classification

We train a Spa-Net to identify the correct pairings and perform the signal/background classification at the same time.

6.1 Training dataset

The selection and matching process for the jet pairing is the same as Section 2.1. We prepare the signal and background samples of the same size for classification.

For the jet assignment part,

- Training sample:
  - Total sample size: 1,800,000
  - 1h sample size: 318,053
  - 2h sample size: 277,876
  - 3h sample size: 162,444
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - 1h sample size: 35,372
  - 2h sample size: 30,853
  - 3h sample size: 18,004

For event classification,

- Training sample:

  - Total sample size: 1,800,000
  - Signal sample size: 900,000
  - Background sample size: 900,000
  - 5% used on validation

- Testing sample:
  - Total sample size: 200,000
  - Signal sample size: 100,000
  - Background sample size: 100,000

This training takes around 10 hours on our server.

6.2 Training results

The training results are presented in Table 7. This result is better than Table 5 since we use larger training datasets.

Table 7: Spa-Net training results on the tri-Higgs samples. Spa-Net is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.656	0.000	0.082	0.262	0.684
$= 7$	0.436	0.017	0.168	0.379	0.504
≥ 8	0.341	0.018	0.173	0.468	0.411
Total	0.478	0.012	0.142	0.368	0.533

Table 8 presents the classification training results. We use the accuracy (ACC) and the area under the Receiver Operating Characteristic (ROC) curve (AUC) as two metrics.

Table 8: The Spa-Net classification training results with tri-Higgs sample.

	ACC	AUC
Spa-Net	0.822	0.900

6.3 3h training dataset

We only consider 3h events in pairing tasks in this subsection. We prepare the signal and background samples of the same size for classification.

For the jet assignment part,

- Training sample:
  - Total sample size: 1,800,000
  - 1h sample size: 0
  - 2h sample size: 0
  - 3h sample size: 900,000
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - 1h sample size: 0
  - 2h sample size: 0
  - 3h sample size: 100,000

For event classification,

- Training sample:
  - Total sample size: 1,800,000
  - Signal sample size: 900,000
  - Background sample size: 900,000
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - Signal sample size: 100,000
  - Background sample size: 100,000

The training results are presented in Table 9. This result is similar to Table 7.

However, some issues should be resolved when we try to use all 3h events for combining training. The loss values are not reasonable.

Table 9: SPA-NET training results on the tri-Higgs samples, where we only consider 3h events. SPA-NET is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.680	0.000	0.084	0.236	0.708
$= 7$	0.477	0.014	0.150	0.359	0.536
≥ 8	0.311	0.027	0.184	0.477	0.391
Total	0.491	0.014	0.139	0.356	0.547

7 Review the pairing methods

In Refs. [5, 6]: We select the 6 b-tagged jets with the highest transverse momentum. The requirements for the transverse momentum and pseudo-rapidity are applied. We subsequently make use of the observable:

$$\chi^{2,(6)} = \sum_{qr \in J} (m_{qr} - m_h)^2 \tag{4}$$

where $J = \{j_1 j_2, j_3 j_4, j_5 j_6\}$ is the set of all possible 15 pairings of 6 b-tagged jets. Out of all the possible combinations, we pick the one with the smallest value $\chi^{2,(6)}_{\text{min}}$. The pairing of b-jets defining $\chi^{2,(6)}_{\text{min}}$ is our best candidate for the reconstruction of the three Higgs bosons, h. No pairing efficiency is provided.

In Ref. [7]: We select the 6 b-tagged jets with the highest transverse momentum and form pairs in different combinations, with the aim of first reconstructing individual SM-like Higgs bosons, h_1, and subsequently the two scalars h_2 and h_3. To this end, we introduce two observables:

$$\chi^{2,(4)} = \sum_{qr \in I} (m_{qr} - m_h)^2 \tag{5}$$

$$\chi^{2,(6)} = \sum_{qr \in J} (m_{qr} - m_h)^2 \tag{6}$$

where we have defined the sets $I = \{i_1 i_2, i_3 i_4\}$ and $J = \{j_1 j_2, j_3 j_4, j_5 j_6\}$, constructed from different pairings of 4 and 6 b-tagged jets, respectively, and where m_{qr} denotes the invariant mass of the respective pairing, qr. Note that the set I that defines $\chi^{2,(4)}_{\text{min}}$ should be a subset of the arrangement J.

We select the combinations of b-tagged jets entering in I and J based on the minimization of the sum

$$\chi^{2,(4)} + \chi^{2,(6)} \tag{7}$$

We then "identify" candidates for the scalars h_2 and h_3 with the pairing configurations $I_{\min}$ and $J_{\min}$ which minimise $\chi^{2,(4)}$ and $\chi^{2,(6)}$ respectively. Note that this procedure does not guarantee that $I_{\min}$ indeed reconstructs to h_2; in fact, we found this to be the case in about 40% on average for all benchmark samples, being slightly higher than a "blind guess" that would lead to a probability of 1/3.

8 6b requirement

They could utilize stronger kinetic and b-tagging requirements in experiments. However, we do not use the b-tag information for the previous χ^2 method. Thus, we change the preselection condition to make a fair comparison.

For testing samples, we only consider the event that contains at least 6 b-tagged jets. The Higgs Bosons are reconstructed by pairing the six leading b-jets for the χ^2 method. Thus, the best one is chosen from 15 possible pairing combinations.

Table 10 and Table 11 are the pairing performance on 6 b-tagged samples. The χ^2 method performs better than the previous results (Table 6) but still performs worse than the SPA-NET pairing.

Table 10: χ^2 pairing efficiencies on different categories. The χ^2 method only considers the six leading b-jets.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.500	0.000	0.124	0.376	0.541
$= 7$	0.322	0.072	0.185	0.486	0.389
≥ 8	0.238	0.026	0.156	0.579	0.308
Total	0.329	0.013	0.159	0.499	0.391

9 Another pairing method

The pairing algorithm is defined to minimize

$$|m_{h_1} - 120| + |m_{h_2} - 115| + |m_{h_3} - 110| \tag{8}$$

where m_{h_i} is the mass of the i-th Higgs boson candidate (sorted by p_{T}) in units of GeV. The numbers in this definition are chosen based on the peaks of the m_{h_i} distributions in simulated signal events.

Table 11: SPA-NET training results on the tri-Higgs samples, where we only consider 3h events. SPA-NET is trained on $\geq 0b$ datasets and tested on $\geq 6b$ datasets.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.679	0.000	0.074	0.246	0.704
$= 7$	0.522	0.013	0.127	0.337	0.574
≥ 8	0.354	0.019	0.180	0.447	0.427
Total	0.492	0.013	0.136	0.359	0.546

Figure 1 is the Higgs boson invariant mass m_h distributions. The peaks of the m_h distributions in our simulation are $121, 119, 116$ GeV. Based on these values, we modify the numbers in Equation 8. Table 12 is the pairing performance on 6 b-tagged samples. The absolute value method performs worse than the χ^2 method (Table 10).

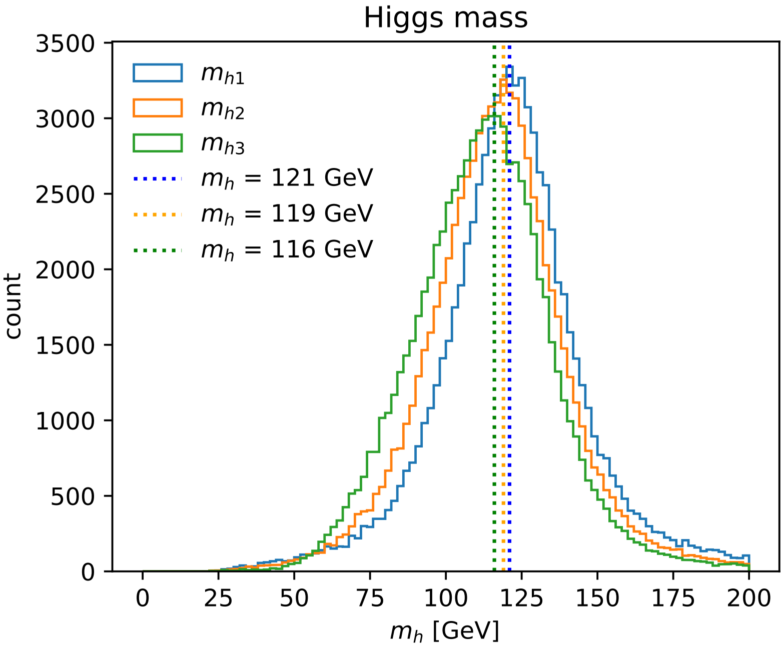


Figure 1: The Higgs boson mass distributions. Here, we only consider the event containing at least 6 b-tagged jets.

Table 12: χ^2 pairing efficiencies on different categories. The χ^2 method only considers the six leading b-jets.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.476	0.000	0.135	0.388	0.522
$= 7$	0.293	0.011	0.185	0.511	0.362
≥ 8	0.215	0.026	0.168	0.589	0.289
Total	0.303	0.015	0.167	0.515	0.367

10 New simulation setting

We apply the new setting to the event generation for a fair comparison. The following are the notes on the latest simulation setting.

The $6b$ analysis uses $R = 0.4$ anti-k_{T} jets. All jets are required to have $p_{\text{T}} > 20$ GeV and $|\eta| < 2.5$, to be within the tracker acceptance for b-tagging.

All events must pass the Preselection. They must have at least six jets, defined by the earlier selection criteria. Of these six jets, at least four jets must have $p_{\text{T}} > 40$ GeV, and at least four jets must be b-tagged.

Pairing is performed on all events passing the Preselection. In the case of $4b$ or $5b$ events, the six jets considered for pairing into Higgs candidates are the b-tagged jets, and the extra jets are selected as the highest p_{T} of the remaining light-flavor jets. In the case of $6b$ events, the 6 jets are the 6 leading p_{T} b-tagged jets.

11 Pairing for new simulated samples

We modified the `Delphes` card to apply the anti-k_{T} clustering algorithm with $R = 0.4$ and change the b-tagging efficiency to the ATLAS DL1r 77% working point. At this working point, the light-jet (charm-jet) rejection is about 130 (4.9).

Figure 2 is the Higgs boson invariant mass m_h distributions for new simulated samples. All events passing the Preselection and matching are used to generate this plot. The peaks of the m_h distributions in our simulation are $119, 115, 111$ GeV. We modify the numbers in Equation 8 based on these values.

Table 13, 14 and 15 are the pairing performance on new simulated samples. The SPA-NET performs the best.

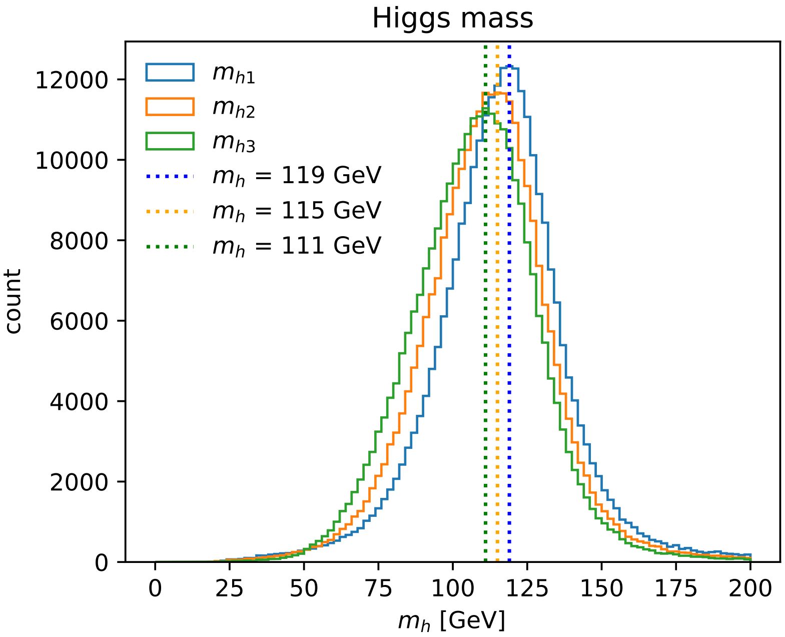


Figure 2: The Higgs boson mass distributions. Here, we consider all events passing the Preselection, and the correct pairing can be obtained.

Table 13: χ^2 pairing efficiencies on different categories. We minimize the quantity defined in Equation 3. The χ^2 method considers the possible combinations of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.241	0.519	0.000	0.112	0.368	0.557
$= 7$	0.335	0.272	0.010	0.186	0.532	0.341
≥ 8	0.424	0.147	0.008	0.196	0.649	0.218
Total	1.000	0.279	0.007	0.172	0.542	0.341

Table 14: Absolute value pairing efficiencies on different categories. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.241	0.457	0.000	0.131	0.412	0.500
$= 7$	0.335	0.239	0.009	0.184	0.568	0.306
≥ 8	0.424	0.136	0.007	0.197	0.661	0.206
Total	1.000	0.248	0.006	0.176	0.570	0.311

Table 15: SPA-NET training results on the tri-Higgs samples, where we only consider 3h events. The SPA-NET method considers all jets in the final state.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.241	0.619	0.000	0.097	0.283	0.652
$= 7$	0.335	0.481	0.008	0.150	0.361	0.536
≥ 8	0.424	0.350	0.011	0.171	0.469	0.414
Total	1.000	0.459	0.007	0.146	0.388	0.512

12 Modify the matching strategy

For the previous exercise, we define the correct jet assignments for each event by matching the jets to the simulated truth quarks within an angular distance of $\Delta R < 0.4$. Such an event will be dropped if a simulated truth quark is matched to more than one jet. Furthermore, some simulated truth quarks may not be matched to any jet, so the event will not be used in training either.

We modify our strategy for more than one jet case. If more than one jet can be matched to a simulated truth quark in the $\Delta R = 0.4$ cone, we choose the nearest one by the ΔR distance. This method is the same as the di-Higgs analysis [8].

Table 16 is the cutflow number at different selection cuts.

Table 16: The number of passing events, efficiencies, and passing rates for signal processes at different selection cuts.

	Count	Efficiency	Pass rate
Total	100000	1.00	1.00
≥ 6 jets	61454	0.61	0.61
≥ 4 jets with $p_{\text{T}} > 40$ GeV	50341	0.82	0.50
≥ 4 b-jets	32337	0.64	0.32
Matching 3h	9944	0.31	0.10
Matching 2h	11341	0.35	0.11
Matching 1h	8941	0.28	0.09

Table 17 and 18 are the pairing performance with the new matching strategy. The results are similar to the previous one (Table 13 and 14).

Table 17: χ^2 pairing efficiencies on different categories. We minimize the quantity defined in Equation 3. The χ^2 method considers the possible combinations of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.242	0.520	0.000	0.112	0.368	0.557
$= 7$	0.335	0.277	0.007	0.182	0.534	0.343
≥ 8	0.422	0.161	0.008	0.190	0.641	0.229
Total	1.000	0.287	0.006	0.168	0.539	0.347

Table 18: Absolute value pairing efficiencies on different categories. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.242	0.450	0.000	0.132	0.418	0.494
$= 7$	0.335	0.249	0.007	0.177	0.566	0.313
≥ 8	0.422	0.145	0.009	0.194	0.653	0.215
Total	1.000	0.254	0.006	0.173	0.567	0.316

13 SM tri-Higgs samples

We prepare the SM tri-Higgs samples to identify the reason for the matching issue.

The SM tri-Higgs process is generated at the centre-of-mass energy $\sqrt{s} = 13$ TeV with the `NNPDF30_nlo_as_0119` PDF set [9]. In `Delphes`, we use the anti-k_T clustering algorithm with $R = 0.4$ and set the b-tagging efficiency to the ATLAS DL1r 77% working point. At this working point, the light-jet (charm-jet) rejection is about 130 (4.9). Following are the MadGraph scripts for generating SM tri-Higgs samples:

```
generate p p > h h h [QCD] QED^2<=6
output MG5/pphhh_sm
launch MG5/pphhh_sm

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

Cards/delphes_card.dat

set run_card nevents 100000
set run_card ebeam1 6500.0
set run_card ebeam2 6500.0

set run_card pdlabel lhapdf
set run_card lhaid 266000

set run_card ptb 19
set run_card etab 2.6

set spinmode none
decay h > b b~

done
```

Table 19 is the cutflow number of the SM tri-Higgs process. The matching efficiency is similar to the resonant tri-Higgs case (Table 16).

Table 19: The number of passing events, efficiencies, and passing rates for signal processes at different selection cuts.

	Count	Efficiency	Pass rate
Total	100000	1.00	1.00
≥ 6 jets	74878	0.75	0.75
≥ 4 jets with $p_{\mathrm{T}} > 40$ GeV	70399	0.94	0.70
≥ 4 b-jets	48734	0.69	0.49
Matching 3h	17599	0.36	0.18
Matching 2h	18601	0.38	0.19
Matching 1h	10759	0.22	0.10

Table 20 and 21 are the pairing performance of the SM tri-Higgs process.

Table 20: χ^2 pairing efficiencies of the SM tri-Higgs samples. We minimize the quantity defined in Equation 3. The χ^2 method considers the possible combinations of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.169	0.556	0.000	0.107	0.338	0.591
$= 7$	0.311	0.334	0.008	0.154	0.505	0.390
≥ 8	0.520	0.193	0.007	0.204	0.596	0.266
Total	1.000	0.298	0.006	0.172	0.524	0.359

14 Use Pythia for Higgs decay

We use `MadSpin` to implement the Higgs decay in the previous exercise, while `Pythia` could also do it. Thus, we generate the samples with `Pythia` decay and compute the cutflow table.

Table 22 is the cutflow number of the Pythia decayed resonant tri-Higgs process. The matching efficiency is similar to the `MadSpin` case (Table 16).

15 Matching rate in different categories

Based on the b-jet multiplicity, we can categorize events into $4b$, $5b$, and $6b$ regions after Preselection. They are required to have exactly 4, exactly 5, or ≥ 6 b-tagged jets,

Table 21: Absolute value pairing efficiencies of the SM tri-Higgs samples. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.169	0.651	0.000	0.095	0.254	0.683
$= 7$	0.311	0.401	0.014	0.158	0.427	0.463
≥ 8	0.520	0.226	0.009	0.213	0.552	0.303
Total	1.000	0.352	0.009	0.176	0.463	0.417

Table 22: The number of passing events, efficiencies, and passing rates for signal processes at different selection cuts.

	Count	Efficiency	Pass rate
Total	10000	1.00	1.00
≥ 6 jets	5209	0.52	0.52
≥ 4 jets with $p_{\mathrm{T}} > 40$ GeV	4501	0.78	0.41
≥ 4 b-jets	2035	0.50	0.20
Matching 3h	705	0.35	0.07
Matching 2h	644	0.32	0.06
Matching 1h	575	0.28	0.06

respectively. We compute the matching efficiency and event fraction in each category to obtain more details about the samples. Similarly, we can compute the matching efficiency and event fraction in different N_{Jet} categories.

Table 23 and 24 are the matching rates in different $N_{b\text{-Jet}}$ and N_{Jet} categories for resonant samples, respectively. Only the events passing the Preselection would be considered.

The matching rate in the $6b$ region is much higher than in the $4b$ and $5b$ regions. However, the event fraction of $6b$ categories is 12%. Therefore, the total matching efficiency is 31%. The matching rate in the $8j$ region is higher than in the $6j$ and $7j$ regions. If there is a jet not decaying from the b-parton, then the matching would fail in the $6j$ case. Thus, this result is satisfied our expectation.

Table 23: The matching rates on different $N_{b\text{-Jet}}$ categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection.

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.509	0.187
$= 5$	0.368	0.354
≥ 6	0.123	0.669
Total	1.000	0.308

Table 24: The matching rates on different N_{Jet} categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection.

N_{Jet}	Fraction	Match Efficiency
$= 6$	0.341	0.218
$= 7$	0.320	0.322
≥ 8	0.338	0.384
Total	1.000	0.308

Table 25 and 26 are the matching rates in different $N_{b\text{-Jet}}$ and N_{Jet} categories for SM samples, respectively.

Similarly, the matching rate in the $6b$ region is much higher than in the $4b$ and $5b$ regions. However, the event fraction of the $6b$ category is much lower than $4b$ and $5b$ regions. Thus, the total matching efficiency is 36%.

Table 25: The matching rates on different $N_{b\text{-Jet}}$ categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection.

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.475	0.227
$= 5$	0.378	0.395
≥ 6	0.147	0.704
Total	1.000	0.361

Table 26: The matching rates on different N_{Jet} categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection.

N_{Jet}	Fraction	Match Efficiency
$= 6$	0.282	0.227
$= 7$	0.307	0.352
≥ 8	0.411	0.460
Total	1.000	0.361

16 Pairing efficiency in different categories

Similarly, we can categorize events into $4b$, $5b$, and $6b$ regions after Preselection, which are required to have exactly 4, exactly 5, or ≥ 6 b-tagged jets, respectively. Then, we compute the pairing efficiency and event fraction in each category to obtain more details about the samples.

Table 27 is the pairing efficiency in different $N_{b\text{-Jet}}$ categories for resonant samples. Only the events passing the Preselection and whose b-partons all can be matched would be considered. The pairing efficiency in the $6b$ region is higher than in the $4b$ and $5b$ regions. However, the event fraction of the $6b$ category only contributes 27%. Therefore, the total pairing efficiency is 25%. Note that these results are consistent with Table 18.

Table 28 is the pairing efficiency in different $N_{b\text{-Jet}}$ categories for SM samples. Similarly, the pairing efficiency in the $6b$ region is higher than in the $4b$ and $5b$ regions. However, the event fraction of the $6b$ category is lower than the $4b$ and $5b$ regions. Thus, the total matching efficiency is 35%. These results are consistent with Table 21.

Table 27: Absolute value pairing efficiencies of the resonant tri-Higgs samples on different $N_{b\text{-Jet}}$ categories. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

		Correctly reconstructed Higgs				
$N_{b\text{-Jet}}$	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 4$	0.309	0.190	0.008	0.189	0.614	0.258
$= 5$	0.424	0.225	0.005	0.179	0.591	0.288
≥ 6	0.267	0.373	0.005	0.146	0.475	0.425
Total	1.000	0.254	0.006	0.173	0.567	0.316

Table 28: Absolute value pairing efficiencies of the SM tri-Higgs samples on different $N_{b\text{-Jet}}$ categories. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

		Correctly reconstructed Higgs				
$N_{b\text{-Jet}}$	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 4$	0.301	0.295	0.011	0.181	0.514	0.362
$= 5$	0.419	0.310	0.011	0.193	0.486	0.381
≥ 6	0.280	0.477	0.005	0.145	0.373	0.529
Total	1.000	0.352	0.009	0.176	0.463	0.417

17 Matching rate in more categories

Similarly to Section 15, we compute the matching efficiency and event fraction in more categories to obtain more details about the samples.

Table 29 is the matching rates in different $N_{b\text{-Jet}}$ categories for resonant samples. Only the events passing the Preselection would be considered. The matching rate in the $6b$ region is similar to the $7b$ and $8b$ regions. However, the event fraction of the $6b$ category is much higher than the $7b$ and $8b$ regions. Therefore, the matching efficiency of the $\geq 6b$ category is dominated by the $6b$ region.

Table 29: The matching rates on different $N_{b\text{-Jet}}$ categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection. No event belongs to the $\geq 9b$ category.

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.509	0.187
$= 5$	0.368	0.354
$= 6$	0.113	0.669
$= 7$	0.009	0.674
$= 8$	0.001	0.650
≥ 9	0.000	nan
Total	1.000	0.308

Table 30 is the matching rates in different $N_{b\text{-Jet}}$ categories for SM samples. Similarly, the matching rate in the $6b$ region is similar to the $7b$ and $8b$ regions. However, the event fraction of $6b$ categories is much higher than the $7b \sim 10b$ regions. Therefore, the matching efficiency of the $\geq 6b$ category is dominated by the $6b$ region.

18 High resonant samples

We generate samples with high resonant mass. Here we choose three benchmarks: $(m_{h_3}, m_{h_2}) = (1500 \text{ GeV}, 1000 \text{ GeV}), (1500 \text{ GeV}, 850 \text{ GeV}), (1300 \text{ GeV}, 1000 \text{ GeV})$.

Table 31 is the matching rates in different $N_{b\text{-Jet}}$ categories for resonant samples. Only the events passing the Preselection would be considered. These three benchmark points have similar match rates in the $6b$ region. This value is also similar to the previous one (Table 23).

Table 30: The matching rates on different $N_{b\text{-Jet}}$ categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection. No event belongs to the $\geq 11b$ category.

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.475	0.227
$= 5$	0.378	0.395
$= 6$	0.132	0.702
$= 7$	0.014	0.723
$= 8$	0.001	0.742
$= 9$	0.000	1.000
$= 10$	0.000	1.000
≥ 11	0.000	nan
Total	1.000	0.361

Table 31: The matching rates on different $N_{b\text{-Jet}}$ categories. The numerator is the number of events each b-parton can be matched to a jet. The denominator is the number of events passing the Preselection.

(a) $m_{h_3} = 1500$ GeV, $m_{h_2} = 1000$ GeV

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.603	0.260
$= 5$	0.316	0.408
≥ 6	0.082	0.662
Total	1.000	0.340

(b) $m_{h_3} = 1500$ GeV, $m_{h_2} = 850$ GeV

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.588	0.279
$= 5$	0.320	0.466
≥ 6	0.092	0.672
Total	1.000	0.375

(c) $m_{h_3} = 1300$ GeV, $m_{h_2} = 1000$ GeV

$N_{b\text{-Jet}}$	Fraction	Match Efficiency
$= 4$	0.578	0.294
$= 5$	0.324	0.447
≥ 6	0.098	0.654
Total	1.000	0.379

Table 32 is the pairing efficiency in different $N_{b\text{-Jet}}$ categories for resonant samples. Only the events passing the Preselection and whose b-partons all can be matched would be considered. Compared with the low resonant case (Table 27), the high resonant samples have higher pairing efficiencies.

Table 32: Absolute value pairing efficiencies of the resonant tri-Higgs samples on different $N_{b\text{-Jet}}$ categories. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

(a) $m_{h_3} = 1500$ GeV$, m_{h_2} = 1000$ GeV

		Correctly reconstructed Higgs				
$N_{b\text{-Jet}}$	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 4$	0.435	0.467	0.037	0.185	0.310	0.554
$= 5$	0.399	0.568	0.029	0.174	0.229	0.645
≥ 6	0.166	0.671	0.012	0.124	0.193	0.720
Total	1.000	0.541	0.030	0.171	0.258	0.618

(b) $m_{h_3} = 1500$ GeV$, m_{h_2} = 850$ GeV

		Correctly reconstructed Higgs				
$N_{b\text{-Jet}}$	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 4$	0.432	0.474	0.040	0.199	0.288	0.566
$= 5$	0.387	0.584	0.028	0.165	0.223	0.657
≥ 6	0.181	0.678	0.019	0.115	0.189	0.729
Total	1.000	0.553	0.032	0.170	0.245	0.631

(c) $m_{h_3} = 1300$ GeV$, m_{h_2} = 1000$ GeV

		Correctly reconstructed Higgs				
$N_{b\text{-Jet}}$	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 4$	0.447	0.448	0.033	0.157	0.361	0.523
$= 5$	0.379	0.525	0.017	0.124	0.334	0.578
≥ 6	0.174	0.644	0.026	0.131	0.199	0.705
Total	1.000	0.511	0.026	0.140	0.323	0.575

19 Matching failed events

Even when considering the $6b$ category, the matching efficiencies remain around 66%. To better understand this, we must investigate the reasons for the failed matching cases.

Matching algorithm: For each simulated truth quark, we compute the ΔR between the quark and all jets and then match the quark to the jet within $\Delta R = 0.4$ cone. If more than one jet can be matched to a simulated truth quark in the $\Delta R = 0.4$ cone, we choose the nearest one by the ΔR distance. If multiple quarks are matched to the same jet, then the matching is failed.

Here, we only consider the $6b$ category. First, we compute the number of quarks that can be matched to jets, meaning that at least one jet is within the $\Delta R = 0.4$ cone of the quark. Events with $N_{\mathrm{q}} < 6$ are the failed matching cases. Table 33 presents the quark matching results.

For the $N_{\mathrm{q}} = 6$ case, if multiple quarks are matched to the same jet, then the matching fails. Two types of $N_{\mathrm{q}} = 6$ failed cases exist. One is for each quark that has only one jet within the $\Delta R = 0.4$ cone. Another is there is more than one jet within the $\Delta R = 0.4$ cone of quark. Table 34 summarizes the vairous cases in the $N_{\mathrm{q}} = 6$ category.

Table 33: The quark matching table. N_{q} is the number of quarks that can be matched to jets in an event.

(a) $m_{h_3} = 420$ GeV, $m_{h_2} = 280$ GeV

N_{q}	Count	Fraction
3	8	0.002
4	89	0.022
5	871	0.220
6	2998	0.756
Total	3966	1.000

(b) SM

N_{q}	Count	Fraction
3	4	0.001
4	156	0.022
5	1505	0.210
6	5510	0.768
Total	7175	1.000

19.1 ΔR distance

Table 33 shows that only 76% event can match all quarks to at least one jet. We modify the ΔR requirement to study more details about matching failed events.

Table 35 presents the quark matching results, where $\Delta R = 0.5$. Increasing the ΔR distance increases the fraction of event with $N_{\mathrm{q}} = 6$ while the values are still lower than

Table 34: The various cases of quark matching results in the $N_{\mathrm{q}} = 6$ case.

(a) $m_{h_3} = 420$ GeV, $m_{h_2} = 280$ GeV

$N_{\mathrm{q}} = 6$	Count	Fraction
Successful matching	2655	0.886
Only one matching jet for each quark	285	0.095
More than one jet for some quark	58	0.019
Total	2998	1.000

(b) SM

$N_{\mathrm{q}} = 6$	Count	Fraction
Successful matching	5054	0.917
Only one matching jet for each quark	376	0.068
More than one jet for some quark	80	0.015
Total	5510	1.000

80%.

Table 35: The quark matching table. N_{q} is the number of quarks that can be matched to jets in an event.

(a) $m_{h_3} = 420$ GeV, $m_{h_2} = 280$ GeV

N_{q}	Count	Fraction
3	7	0.002
4	73	0.018
5	768	0.194
6	3118	0.786
Total	3966	1.000

(b) SM

N_{q}	Count	Fraction
3	2	0.000
4	113	0.016
5	1321	0.184
6	5739	0.800
Total	7175	1.000

20 Train SPANet with the samples in $6b$ region

In this section, we only consider the signal events in the $6b$ region, i.e., there ≥ 6 b-tagged jets. We prepare the samples in the $6b$ region for the Spa-Net training.

For the jet assignment part,

- Training sample:
    - Total sample size: 360,000
    - 1h sample size: 38,595
    - 2h sample size: 73,036
    - 3h sample size: 243,762
    - 5% used on validation
- Testing sample:
    - Total sample size: 40,000
    - 1h sample size: 4,360
    - 2h sample size: 8,070
    - 3h sample size: 27,087

Table 36, 37 and 38 are the pairing performance on the $6b$ samples. The Spa-Net performs the worst. One reason is the sample is not enough for training. We need to generate more training samples to improve the training performance.

Table 36: χ^2 pairing efficiencies of the resonant tri-Higgs samples on different categories. The testing set is the $6b$ samples. We minimize the quantity defined in Equation 3. The χ^2 method considers the possible combinations of 6 highest p_T b-jets.

		Correctly reconstructed Higgs				
N_Jet	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.211	0.509	0.000	0.363	0.127	0.630
$= 7$	0.318	0.429	0.024	0.371	0.176	0.569
≥ 8	0.471	0.311	0.060	0.376	0.254	0.476
Total	1.000	0.390	0.036	0.371	0.203	0.538

Another test is that we prepare the training dataset using different selection criteria. We only require at least 4 b-jets for this dataset. Thus, the sample sizes are the following:

- Training sample:
    - Total sample size: 900,000
    - 1h sample size: 246,462

Table 37: Absolute value pairing efficiencies of the resonant tri-Higgs samples on different categories. The testing set is the $6b$ samples. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 highest p_{T} b-jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.211	0.455	0.000	0.375	0.170	0.580
$= 7$	0.318	0.388	0.024	0.384	0.204	0.532
≥ 8	0.471	0.278	0.057	0.387	0.278	0.445
Total	1.000	0.350	0.035	0.384	0.231	0.501

Table 38: The pairing performance of SPA-NET trained on $6b$ datasets. The testing set is the $6b$ samples. The SPA-NET method considers all jets in the final state.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.211	0.417	0.000	0.414	0.168	0.555
$= 7$	0.318	0.318	0.024	0.440	0.218	0.480
≥ 8	0.471	0.219	0.058	0.425	0.298	0.400
Total	1.000	0.292	0.035	0.427	0.245	0.458

  - 2h sample size: 318,057
  - 3h sample size: 280,788
  - 5% used on validation

- Testing sample:
  - Total sample size: 100,000
  - 1h sample size: 27,243
  - 2h sample size: 35,050
  - 3h sample size: 31,499

Table 39 is the pairing performance of Spa-Net training on the $4b$ samples and testing on $6b$ events. The Spa-Net performs better than the previous one (Table 38). Even if we only test on $6b$ samples, the event with 4 or 5 b-jets can improve the training performance.

Table 39: The pairing performance of Spa-Net trained on the datasets with the at least 4 b-jets requirement. The testing set is the $6b$ samples. The Spa-Net method considers all jets in the final state.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.211	0.648	0.000	0.275	0.077	0.740
$= 7$	0.318	0.539	0.039	0.308	0.114	0.668
≥ 8	0.471	0.390	0.082	0.350	0.177	0.562
Total	1.000	0.492	0.051	0.321	0.136	0.633

21 DNN classifier

To distinguish the signal and background events, we use deep neural networks (DNNs) to classify events as signal-like or background-like.

21.1 Input features

The following are the input features:

- $\Delta R_{h_1}, \Delta R_{h_2}, \Delta R_{h_3}$: The angular distance between the 2 jets form the leading, subleading and least-leading Higgs Boson candidate, respectively.

- RMS $\Delta R_{\mathrm{dijet}}$: The root mean square of the angular distance between all possible dijet combinations that can form a Higgs Boson candidate. We consider all possible permutations of 6 jets selected for pairing.

- Skewness $\Delta A_{\mathrm{dijet}}$: The skewness of $\cosh(\Delta\eta_{ij}) - \cos(\Delta\phi_{ij})$, where i, j are all possible dijet combinations that can form a Higgs Boson candidate.

- $H_{\mathrm{T,\ 6jets}}$: The scalar sum of the p_{T} of the 6 jets that can form 3 Higgs Boson candidates.

- $m_h \cos\theta$: θ is the angle between the reconstructed and reference mass vectors. The reconstructed mass vector is defined as:

$$(m_{h_1}, m_{h_2}, m_{h_3}) - (120, 115, 110)\,\mathrm{GeV}, \tag{9}$$

  where $m_{h_1}, m_{h_2}, m_{h_3}$ are the invariant mass of Higgs candidates. The reference mass vector is formed by the origin to $(120, 115, 110)$ GeV.

- $\eta - m_{hhh}$ fraction: Defined as

$$\frac{\sum_{i,j} 2p_{\mathrm{T},i}p_{\mathrm{T},j}\left(\cosh(\Delta\eta_{ij}) - 1\right)}{m_{hhh}^2} \tag{10}$$

  where i, j are all possible dijet that can form a Higgs Boson candidate and m_{hhh} is the reconstructed tri-Higgs invariant mass.

- Shpericity and Aplanarity of 6 jets. Compute the momentum tensor M_{xyz}

$$M_{xyz} = \sum_i \begin{pmatrix} p_{xi}^2 & p_{xi}p_{yi} & p_{xi}p_{zi} \\ p_{xi}p_{yi} & p_{yi}^2 & p_{yi}p_{zi} \\ p_{xi}p_{zi} & p_{yi}p_{zi} & p_{zi}^2 \end{pmatrix} / \sum_i |p_i|^2 \tag{11}$$

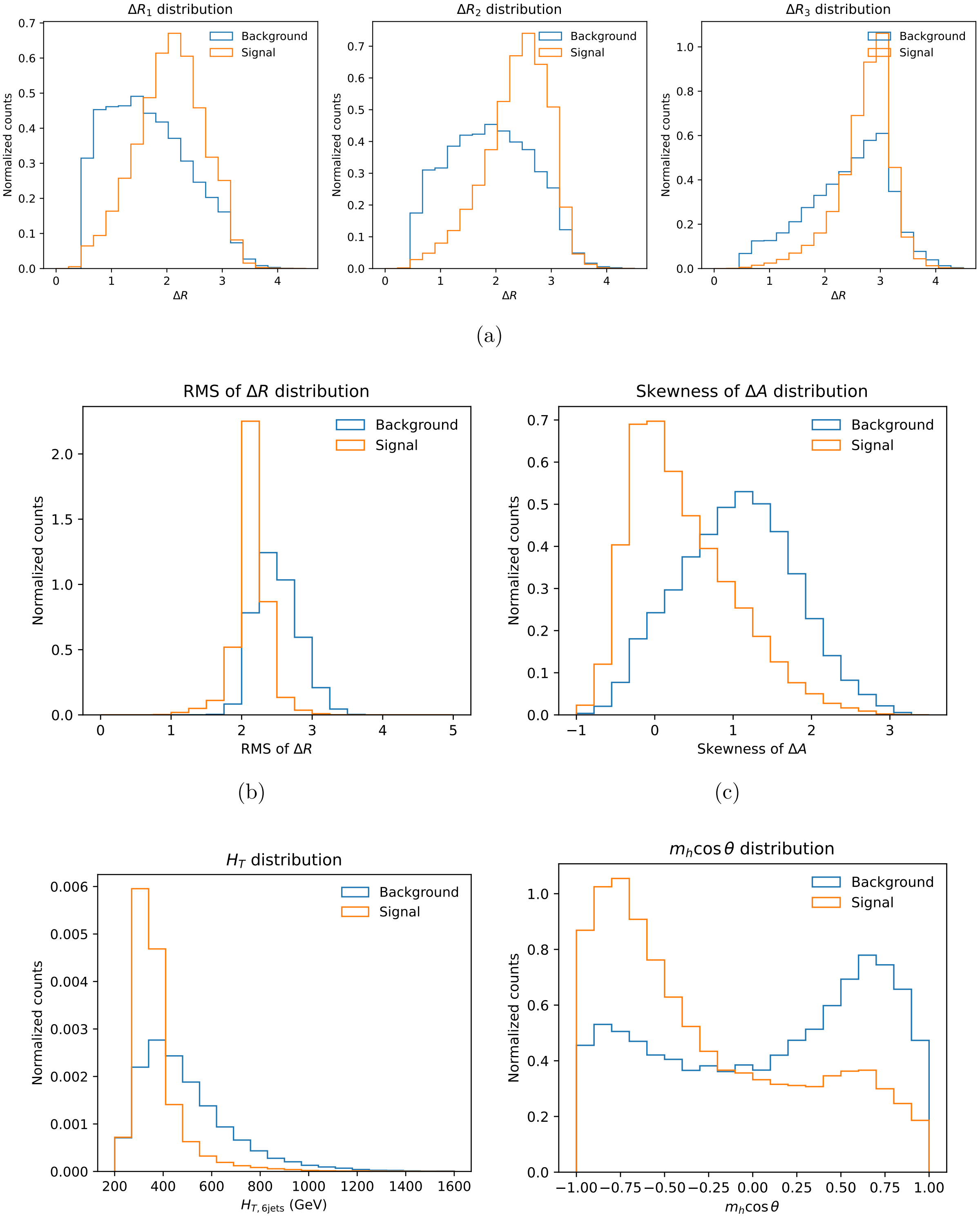
  where i is the jet index. Its eigenvalues are ordered such that $\lambda_1 > \lambda_2 > \lambda_3$. Sphericity is defined as

$$S = \frac{3}{2}(\lambda_2 + \lambda_3) \tag{12}$$

  The aplanarity is defined as

$$A = \frac{3}{2}\lambda_3 \tag{13}$$

The signal is the resonant sample described in Section 1. The background is the $pp \to 6b$ events. We utilize the events in the $6b$ category to plot the input feature distributions. Some variables depend on the pairing results, for example, $\Delta R_{h_i}, m_h \cos\theta$. Strictly speaking, other variables would also depend on how to select the 6 jets for pairing. Figure 3 and 4



(a)

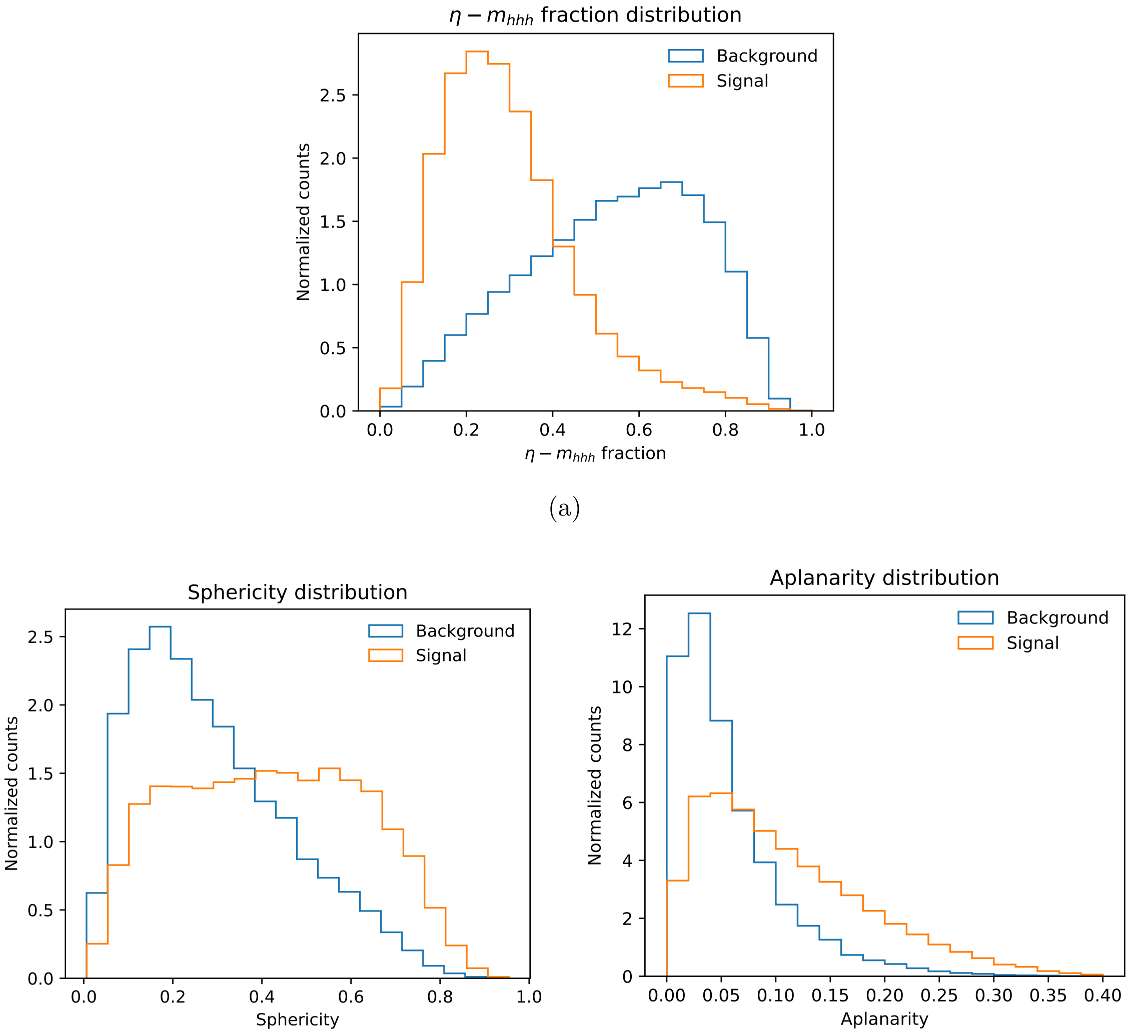
(b)

(c)

(d)

(e)

Figure 3: Distributions of the DNN input features.



(a)

(b)

(c)

Figure 4: Distributions of the DNN input features.

are the input feature distributions. We employ the absolute value pairing to construct these variables.

Figure 5 is the input feature distributions of various pairing methods. Since the distribution of ΔR_{h_i} and $m_h \cos\theta$ depend on the pairing results, we show the plots of these two variables. For other variables, the distributions only depend on how to select the 6 jets for pairing. The distribution of various pairing methods looks similar but is not the same.

21.2 Model structure

We use the fully connected network with the ReLU activation function. After each dense layer, we add a dropout layer. The binary cross entropy is used as the loss function, and the Adam optimizer is used to minimize the loss value. We utilize the early stopping technique to prevent over-training.

Table 40 shows the hyperparameters used in this exercise.

Table 40: The hyperparameter sets of the dense neural network training.

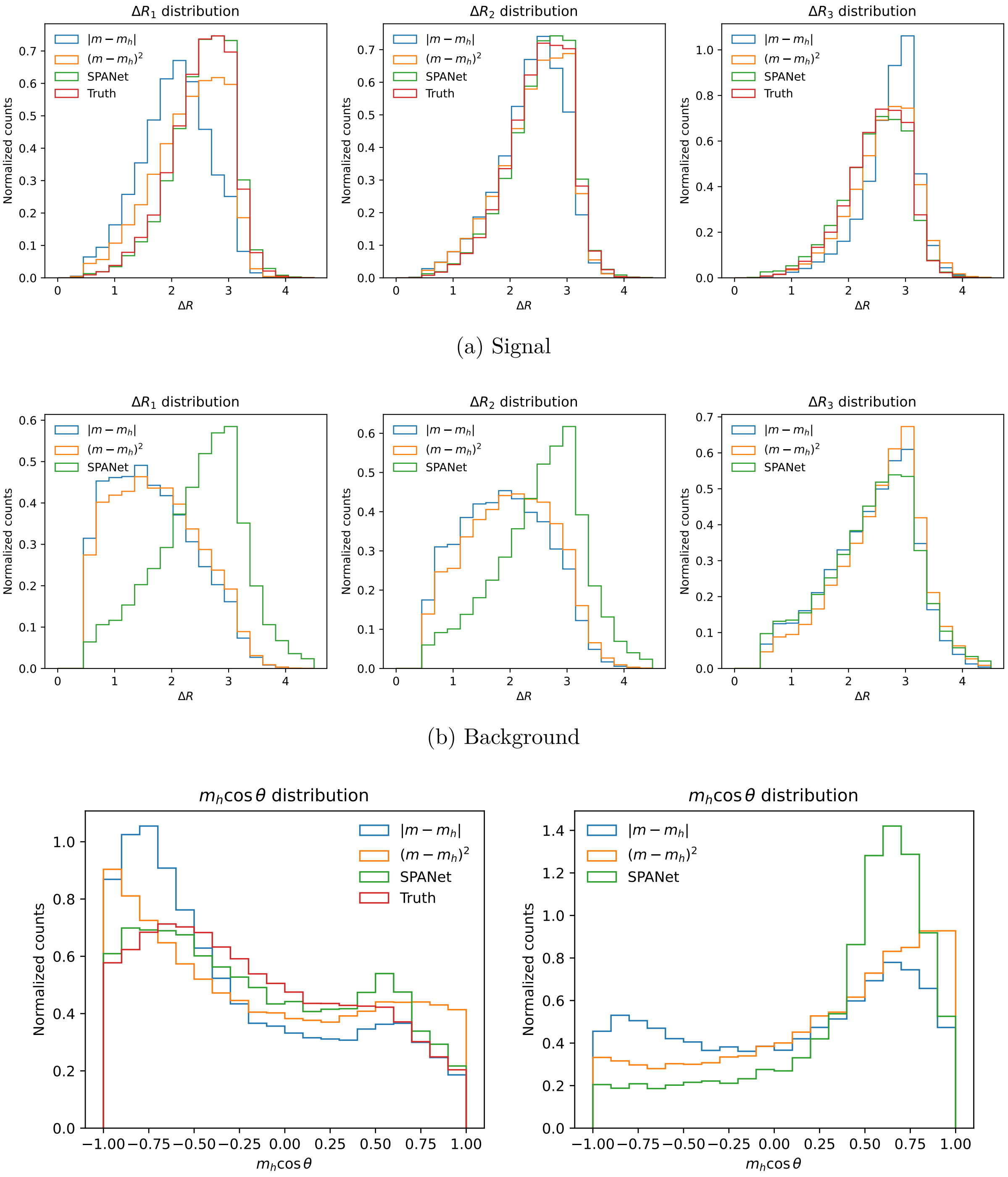
Parameter	Value
Learning rate	0.001
Batchsize	128
Number of hidden layers	3
Number of nodes in each layer	24
Dropout rate	0.1
Patience	10

21.3 Training performance

We use 25k signal and 25k background events, where 70% for training, 15% for validation, and 15% for testing. We use various pairing methods to construct the input features.

Figure 6 illustrates the loss and accuracy across the training process. Figure 7 shows the event score distribution and the ROC curve. The neural network is evaluated on the testing samples to generate these plots. This neural network is trained on the dataset generated from the absolute value pairing method.

Table 41 summarizes the training results of various pairing methods. All methods exhibit similar performance. This suggests that the differences in the input feature distributions are not significant enough to influence the training results.



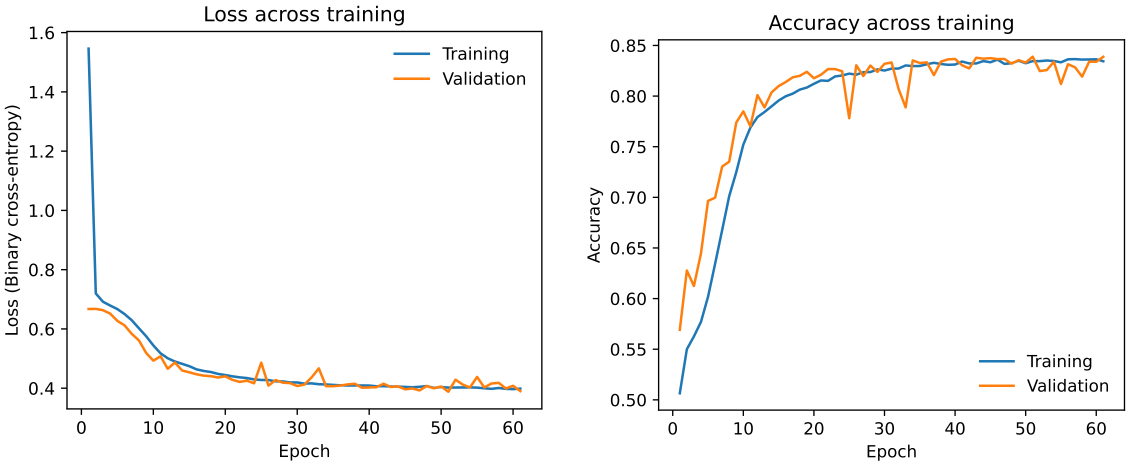
(a) Signal

(b) Background

(c) Signal

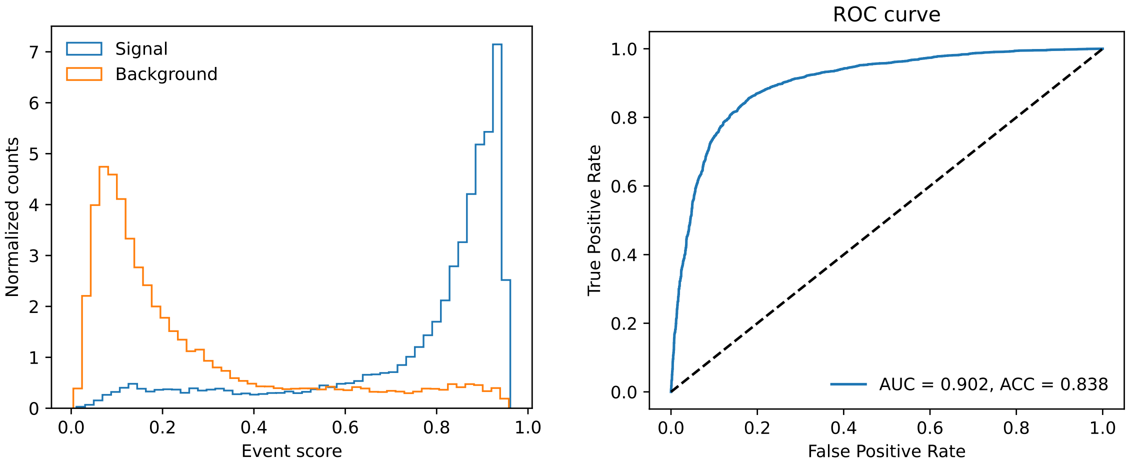
(d) Background

Figure 5: Distributions of ΔR_{h_i} and $m_h \cos\theta$. $|m - m_h|$ is the absolute value pairing method. $(m - m_h)^2$ is the χ^2 pairing method. SPANet is the pairing results from the $4b$ Spa-Net. Truth is the result of the truth pairing.



(a) (b)

Figure 6: The training and validation results during the training process.



(a) (b)

Figure 7: (a) The event score distribution of the testing dataset. (b) The ROC curve of the testing dataset. The AUC is the area under the ROC curve, and the ACC is the best accuracy.

Table 41: The dense neural network training results. The ACC and AUC are evaluated based on 10 trainings.

Pairing method	ACC	AUC
Absolute value	0.836 ± 0.003	0.899 ± 0.005
χ^2	0.836 ± 0.002	0.902 ± 0.003
SPA-NET	0.831 ± 0.006	0.900 ± 0.007

21.4 Training performance with larger datasets

We use 50k signal and 50k background events, where 70% for training, 15% for validation, and 15% for testing. We use various pairing methods to construct the input features.

Table 42 summarizes the training results of various pairing methods. Even if we enlarge the training datasets, all methods exhibit similar performance. This suggests that the differences in the input feature distributions are not significant enough to influence the training results.

Table 42: The dense neural network training results. The ACC and AUC are evaluated based on 10 trainings. 25k means that we use 25k signal and 25k background events.

Sample size	Pairing method	ACC	AUC
25k	Absolute value	0.836 ± 0.003	0.899 ± 0.005
	χ^2	0.836 ± 0.002	0.902 ± 0.003
	Spa-Net	0.831 ± 0.006	0.900 ± 0.007
50k	Absolute value	0.846 ± 0.002	0.910 ± 0.003
	χ^2	0.842 ± 0.004	0.907 ± 0.004
	Spa-Net	0.837 ± 0.005	0.907 ± 0.005

21.5 4b dataset

We only consider the events with 6 b-tagged jets in previous sections. However, only a few events can pass this requirement. To enlarge the training datasets, we lose the requirement to 4 b-tagged jets. The amount of $4b$ datasets is 10 times larger than $6b$ datasets.

Table 43 summarizes the training results of $4b$ datasets. All methods exhibit similar performance. These results are consistent with previous results.

22 SPANet classifier

Spa-Net can be treated as a classifier. In this section, we train a Spa-Net to identify the correct pairings and perform the signal/background classification simultaneously.

22.1 Training dataset

The selection and matching process for the jet pairing is described in section 10 and 12. We only require the 4 b-tagged jets on training datasets. We prepare the signal and

Table 43: The dense neural network training results. The ACC and AUC are evaluated based on 10 trainings. 50k means that we use 50k signal and 50k background events.

		Test on $4b$ datasets		Test on $6b$ datasets	
Sample size	Pairing method	ACC	AUC	ACC	AUC
50k	Absolute value	0.782 ± 0.057	0.844 ± 0.076	0.827 ± 0.038	0.886 ± 0.048
	χ^2	0.800 ± 0.004	0.867 ± 0.005	0.837 ± 0.003	0.900 ± 0.002
	SPA-NET	0.798 ± 0.004	0.870 ± 0.008	0.829 ± 0.004	0.899 ± 0.006
100k	Absolute value	0.803 ± 0.005	0.870 ± 0.006	0.841 ± 0.003	0.903 ± 0.003
	χ^2	0.803 ± 0.005	0.871 ± 0.005	0.838 ± 0.003	0.901 ± 0.002
	SPA-NET	0.807 ± 0.008	0.878 ± 0.007	0.834 ± 0.006	0.905 ± 0.005

background samples of the same size for classification.

For the jet assignment part,

- Training sample:
  - Total sample size: 900,000
  - 1h sample size: 123,141
  - 2h sample size: 159,219
  - 3h sample size: 140,132
  - 5% used on validation
- Testing sample:
  - Total sample size: 100,000
  - 1h sample size: 13,785
  - 2h sample size: 17,448
  - 3h sample size: 15,668

For event classification,

- Training sample:
  - Total sample size: 900,000
  - Signal sample size: 450,000
  - Background sample size: 450,000

- 5% used on validation

- Testing sample:
  - Total sample size: 100,000
  - Signal sample size: 50,000
  - Background sample size: 50,000

This training takes around 4.5 hours on our server.

22.2 Training results

The training results are presented in Table 44.

Table 44: Spa-Net training results on the $4b$ tri-Higgs samples. Spa-Net is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.656	0.000	0.082	0.262	0.684
$= 7$	0.436	0.017	0.168	0.379	0.504
≥ 8	0.341	0.018	0.173	0.468	0.411
Total	0.478	0.012	0.142	0.368	0.533

Table 45 presents the classification training results. We use the accuracy (ACC) and the area under the Receiver Operating Characteristic (ROC) curve (AUC) as two metrics. The results are much better than the DNN classifiers (Table 43). However, one thing should be noted: the training sample size of Spa-Net is 5 times greater than DNN's.

Table 45: The Spa-Net classification training results with $4b$ tri-Higgs sample.

	Test on $4b$ datasets		Test on $6b$ datasets	
	ACC	AUC	ACC	AUC
Spa-Net	0.935	0.984	0.939	0.986

23 DNN classifier 2

23.1 Training performance with 500k datasets

To make a fair comparison of the Spa-Net classifier, we prepare the 500k $4b$ training dataset. The input feature and hyperparameter setting are similar to what we described in section 21.

Table 46 summarizes the training results of $4b$ datasets. Even though we enlarged the sample size to 500k, all methods exhibit similar performance. The Spa-Net outperforms the dense neural network.

Table 46: The dense neural network training results. The ACC and AUC are evaluated based on 10 trainings. 50k means that we use 50k signal and 50k background events.

		Test on $4b$ datasets		Test on $6b$ datasets	
Sample size	Pairing method	ACC	AUC	ACC	AUC
50k	Absolute value	0.782 ± 0.057	0.844 ± 0.076	0.827 ± 0.038	0.886 ± 0.048
	χ^2	0.800 ± 0.004	0.867 ± 0.005	0.837 ± 0.003	0.900 ± 0.002
	Spa-Net	0.798 ± 0.004	0.870 ± 0.008	0.829 ± 0.004	0.899 ± 0.006
100k	Absolute value	0.803 ± 0.005	0.870 ± 0.006	0.841 ± 0.003	0.903 ± 0.003
	χ^2	0.803 ± 0.005	0.871 ± 0.005	0.838 ± 0.003	0.901 ± 0.002
	Spa-Net	0.807 ± 0.008	0.878 ± 0.007	0.834 ± 0.006	0.905 ± 0.005
500k	Absolute value	0.813 ± 0.007	0.881 ± 0.012	0.844 ± 0.007	0.906 ± 0.008
	χ^2	0.813 ± 0.004	0.882 ± 0.005	0.844 ± 0.003	0.908 ± 0.004
	Spa-Net	0.810 ± 0.004	0.881 ± 0.006	0.836 ± 0.004	0.907 ± 0.004

23.2 Modify hyperparameter

We modify the hyperparameter set to explore its impact on training performance. We increase both the number of hidden nodes and the batch size. Table 47 lists the hyperparameters used in this subsection.

Table 48 summarizes the training results of 500k $4b$ datasets with this updated hyperparameter set. The performance of dense neural networks is better than previous results (Table 46), while the Spa-Net still outperforms the dense neural network.

Table 47: The hyperparameter sets of the dense neural network training.

Parameter	Value
Learning rate	0.001
Batchsize	1024
Number of hidden layers	3
Number of nodes in each layer	256
Dropout rate	0.1
Patience	10

Table 48: The dense neural network training results. The ACC and AUC are evaluated based on 10 trainings.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.8290 ± 0.0002	0.9024 ± 0.0002	0.8552 ± 0.0015	0.9221 ± 0.0006
χ^2	0.8288 ± 0.0002	0.9022 ± 0.0002	0.8537 ± 0.0009	0.9212 ± 0.0006
Spa-Net	0.8318 ± 0.0003	0.9065 ± 0.0001	0.8543 ± 0.0009	0.9241 ± 0.0003

24 Upper limit

The neural network event scores p_{event} are utilized to set the upper limit of the cross-section.

The binned p_{event} distribution is considered. The likelihood function L consisting of a product of Poisson distributions

$$L(\text{data}) = \prod_{i=1}^{B} \text{Pois}(n_i \mid n_{i,\text{exp}}) \tag{14}$$

where B is the number of bins, n_i is the number of events in bin i from data, $n_{i,\text{exp}}$ is the expected number of events in bin i. The expected number of events is the sum of the signal and background events.

The Poisson distribution is expressed as:

$$\text{Pois}(n \mid \lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^n}{n!}. \tag{15}$$

Taking the logarithm yields:

$$\begin{aligned}\ln\left(\prod_{i=1}^{B}\mathrm{Pois}(n_i \mid n_{i,\mathrm{exp}})\right) &= \sum_{i=1}^{B}\ln\left(\frac{\mathrm{e}^{-n_{i,\mathrm{exp}}}n_{i,\mathrm{exp}}^{n_i}}{n_i!}\right)\\ &= \sum_{i=1}^{B}\left[-n_{i,\mathrm{exp}} + n_i\ln(n_{i,\mathrm{exp}}) - \ln(n_i!)\right]\end{aligned} \tag{16}$$

where the term $\ln(n_i!)$ is independent of the signal and can be treated as a constant.

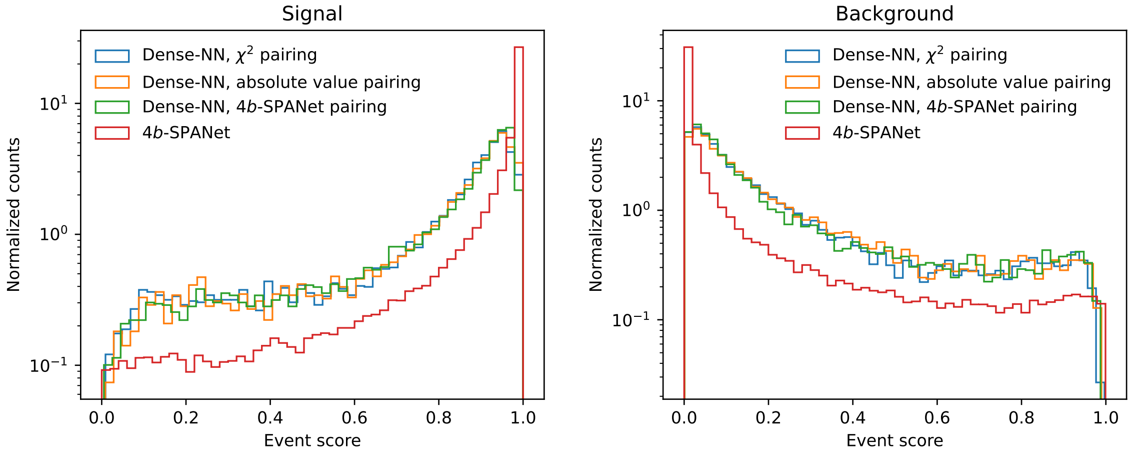
24.1 Event score distribution

Figure 8 shows the p_{event} distributions. The Dense-NN and Spa-Net classifiers were trained as described in sections 23.2 and 22, respectively.

For dense neural networks, the distributions are similar across various pairing methods. These results are consistent with the similar performance shown in table 48. In contrast, the Spa-Net classifier exhibits a different distribution since its accuracy is higher than DNN's by 8%.

Parameter setting:

- Number of bins: 50
- Range: $[0, 1]$



(a) (b)

Figure 8: The p_{event} distribution with various selection methods.

24.2 CLs method

The $\mathrm{CL_s}$ method is used to set the upper limits of the cross-section. The signal strength μ_s is chosen as the parameter of interest (POI). The POI is excluded at the 95% confidence level when the $\mathrm{CL_s} < 0.05$. The package `pyhf` [10, 11] is utilized to calculate the upper limit. Once the upper limit of signal strength is obtained, it can be converted to the upper limit of the cross-section.

Table 49 shows the 95% CL upper limits for μ_s. Dense-NN classifiers with different pairing methods demonstrate similar results. The SPA-NET classifier performs the best among all selection methods. Note that these values are preliminary, as we need to use the correct cross-sections to normalize them. The final results will differ by an overall factor.

Table 49: 95% CL upper limits of signal strength μ_s.

Selection, Pairing	μ_s
Dense-NN, Absolute value	2.530
Dense-NN, χ^2	2.529
Dense-NN, SPA-NET	2.526
SPA-NET	2.185

25 Organize generated samples

All generated samples must pass the pre-selection criteria mentioned in section 10. The samples that meet these criteria are called the $4b$ datasets. Additionally, events containing at least $6b$ jets are used to construct the $6b$ datasets.

The $4b$ datasets consist of 1M signal events and 1M background events, with 90% of the dataset used for training and 10% for testing. These datasets are utilized to train the signal/background classifier.

The $6b$ datasets consist of 50k signal events and 50k background events, and are employed to test event classifiers and determine the upper limits of the cross-section.

25.1 Event classifier

To ensure fair comparisons, both Dense-NN and SPA-NET use the same $4b$ datasets for training. The selection and jet pairing process is described in sections 10 and 12. For training, only the four b-tagged jets are required.

For the jet assignment part (only for SPA-NET), the details of training and testing samples are listed as follows:

- Training sample:
    - Total sample size: 1,800,000
    - 1h sample size: 246,462
    - 2h sample size: 318,057
    - 3h sample size: 280,788
    - 5% used on validation
- Testing sample:
    - Total sample size: 200,000
    - 1h sample size: 27,243
    - 2h sample size: 35,050
    - 3h sample size: 31,499

For event classification, both Dense-NN and SPA-NET

- Training sample:
    - Total sample size: 1,800,000
    - Signal sample size: 900,000
    - Background sample size: 900,000
    - 5% used on validation
- Testing sample:
    - Total sample size: 200,000
    - Signal sample size: 100,000
    - Background sample size: 100,000

Training SPA-NET on our server requires approximately 9.5 hours.

The inputs to the SPA-NET classifiers are the four-momentum vectors of the jets in the final state. The inputs to the Dense-NN classifiers are high-level observables, which are constructed based on the jet pairing methods.

Note that, for the Dense-NN with SPA-NET pairing, we utilized the same $4b$ datasets to train both neural networks for simplicity. Using different datasets for training each network would degrade the pairing performance, thereby increasing the difficulty of classification tasks.

Table 50 summarizes the classification performance of SPA-NET using the $4b$ tri-Higgs dataset. The results are similar to those obtained from the 500k $4b$ dataset, as shown in Table 45.

Table 50: The SPA-NET classification training results with $4b$ tri-Higgs sample.

	Test on $4b$ datasets		Test on $6b$ datasets	
	ACC	AUC	ACC	AUC
SPA-NET	0.928	0.980	0.936	0.984

Table 51 shows the Dense-NN training results for the $4b$ datasets. While the performance of Dense-NN improves with the larger dataset (compared to Table 48), SPA-NET still outperforms Dense-NN.

Table 51: The dense neural network training results. The ACC and AUC are evaluated based on 10 training.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.8418 ± 0.0002	0.9116 ± 0.0001	0.8651 ± 0.0009	0.9317 ± 0.0005
χ^2	0.8412 ± 0.0002	0.9114 ± 0.0001	0.8645 ± 0.0006	0.9311 ± 0.0004
SPA-NET	0.8450 ± 0.0002	0.9164 ± 0.0002	0.8648 ± 0.0004	0.9330 ± 0.0004

25.2 Upper limit constraints

The neural network event scores p_{event} are utilized to set the upper limit of the cross-section. We apply the classifiers mentioned in section 25.1 to the $6b$ datasets to obtain the p_{event} distribution. The details about how to set the upper limits are described in section 24.

Table 52 shows the 95% CL upper limits for μ_{s}. Dense-NN classifiers with different pairing methods demonstrate similar results. The SPA-NET classifier performs the best among all selection methods, which is better than the Dense-NN classifier by about 12%. Note that we need to use the correct cross-sections to normalize them. The final results will differ by an overall factor.

Table 52: 95% CL upper limits of signal strength μ_s.

Selection, Pairing	μ_s
Dense-NN, Absolute value	2.481
Dense-NN, χ^2	2.484
Dense-NN, SPA-NET	2.479
SPA-NET	2.179

26 Different signal benchmarks

For the previous exercise, we always consider $(m_{h_3}, m_{h_2}) = (420, 280)$ GeV. This section generates samples with different resonant masses. Here, we choose three other benchmark points: $(m_{h_3}, m_{h_2}) = (425, 250), (500, 275), (500, 300)$ GeV.

All events must pass the pre-selection criteria. For each benchmark point, we prepare 250k $4b$ events and 12.5k $6b$ events.

However, the MadGraph cannot always generate the required number of events for a mass point $(m_{h_3}, m_{h_2}) = (425, 250)$ GeV. This might result from the $m_{h_2} = 2m_{h_1}$. Thus, we choose another mass point $(m_{h_3}, m_{h_2}) = (520, 325)$ GeV

26.1 Event classifier with mixed datasets

The $4b$ datasets consist of events from all benchmark points. The number of events for each mass point is 250k, and the total size of the $4b$ datasets is 1M.

For the jet assignment part (only for SPA-NET), the details of training and testing samples are listed as follows:

- Training sample:
  - Total sample size: 1,800,000
  - 1h sample size: 246,041
  - 2h sample size: 316,793
  - 3h sample size: 279,125
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - 1h sample size: 27,329

  - 2h sample size: 35,226
  - 3h sample size: 31,100

For event classification, both Dense-NN and SPA-NET

- Training sample:
  - Total sample size: 1,800,000
  - Signal sample size: 900,000
  - Background sample size: 900,000
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - Signal sample size: 100,000
  - Background sample size: 100,000

Training SPA-NET on our server requires approximately 7.6 hours.

The inputs to the SPA-NET classifiers are the four-momentum vectors of the jets in the final state. The inputs to the Dense-NN classifiers are high-level observables, which are constructed based on the jet pairing methods.

Note that, for the Dense-NN with SPA-NET pairing, we utilized the same $4b$ datasets to train both neural networks for simplicity.

Table 53 and 54 summarizes the training performance of SPA-NET using the mixed mass $4b$ tri-Higgs dataset. The training performance is worse than the one considering only one mass point.

Table 55 shows the Dense-NN training results for the mixed mass $4b$ datasets. SPA-NET still outperforms Dense-NN, which is consistent with the previous results.

26.2 Performance at different mass points

We evaluate the pairing and classification performance at each mass point to investigate the detailed results for mixed datasets.

Figure 9 illustrates the pairing performance across different mass points. For the $4b$ dataset, SPA-NET has the highest event efficiency across all mass points. However, for the $6b$ dataset, SPA-NET does not always outperform the cut-based methods. One possible reason is that the training sample size is insufficient.

Table 53: SPA-NET training results on the mixed mass $4b$ tri-Higgs samples. We used the $4b$ testing sample to evaluate the pairing performance. SPA-NET is trained on jet pairing and event classification tasks at the same time.

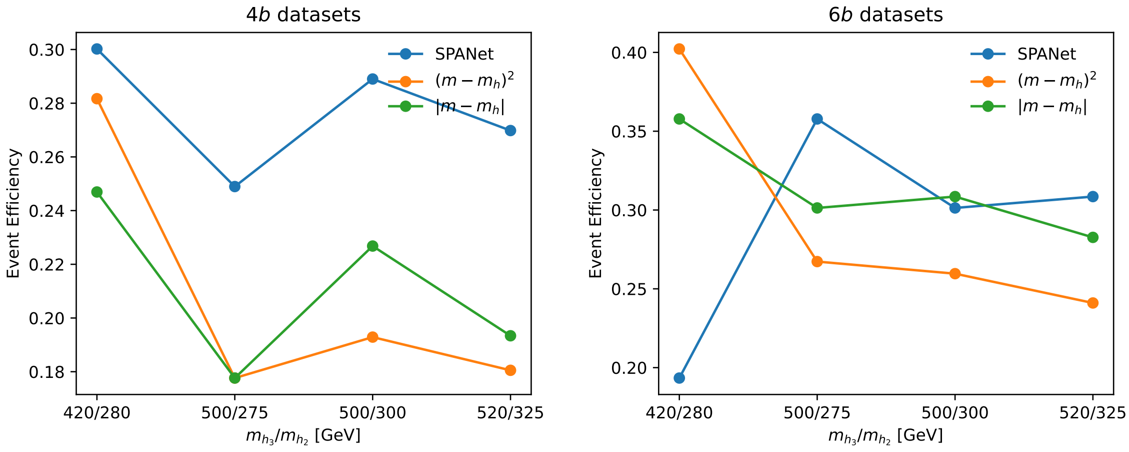
	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.434	0.000	0.401	0.165	0.568
$= 7$	0.289	0.067	0.414	0.230	0.472
≥ 8	0.190	0.100	0.403	0.307	0.391
Total	0.277	0.067	0.406	0.250	0.457

Table 54: The SPA-NET classification training results with mixed mass $4b$ tri-Higgs sample.

	Test on $4b$ datasets		Test on $6b$ datasets	
	ACC	AUC	ACC	AUC
SPA-NET	0.874	0.949	0.905	0.968

Table 55: The dense neural network training results. The ACC and AUC are evaluated based on 10 training.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.7922 ± 0.0003	0.8680 ± 0.0002	0.8262 ± 0.0004	0.8968 ± 0.0003
χ^2	0.7916 ± 0.0002	0.8670 ± 0.0001	0.8248 ± 0.0004	0.8953 ± 0.0003
SPA-NET	0.7921 ± 0.0003	0.8679 ± 0.0001	0.8215 ± 0.0008	0.8939 ± 0.0003



(a)

(b)

Figure 9: Pairing performance at different mass points. We evaluate the event efficiency of various pairing methods. The SPA-NET is trained on the mixed datasets described in Section 26.1.

Figure 10 shows the classification performance at different mass points. The SPA-NET classifier outperforms the Dense-NN classifier for both $4b$ and $6b$ datasets. Dense-NN classifiers with different pairing methods show similar performance trends. These results are consistent with tables 54 and 55.

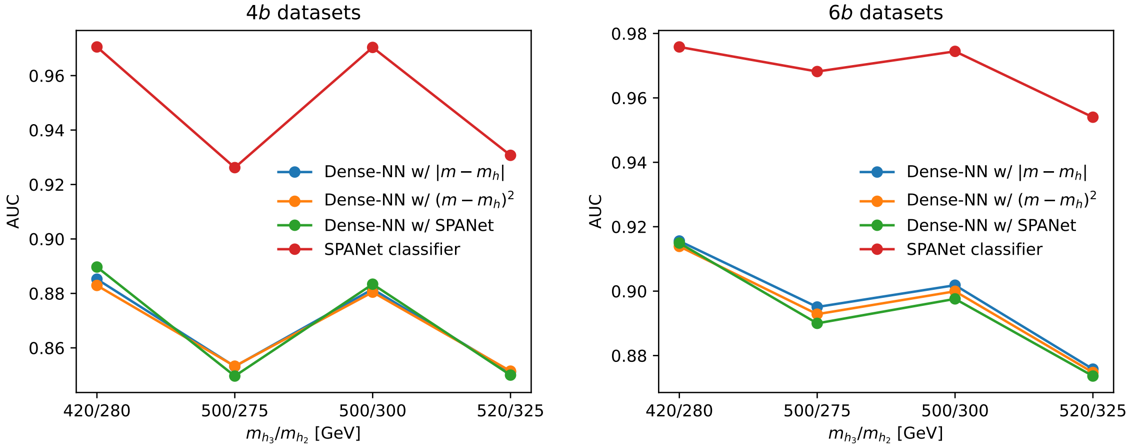
27 New version of SPANet

Previously, I used version 2.0 for SPA-NET training. This section tests version 2.2 and 2.3 (latest version) of SPA-NET. After installing the updated code from GitHub, I trained SPA-NET using $t\bar{t}$ datasets. The results obtained with version 2.2 or 2.3 show a slight improvement over version 2.0, with an accuracy increase of less than 0.5%. The performance difference between the two versions is not significant.

28 More mass points

Here, we choose other benchmark points: $(m_{h_3}, m_{h_2}) = (570,\ 250)$, (600, 325), (700, 325), (800, 325), (700, 400), (800, 400) GeV. All events must pass the pre-selection criteria. We prepare the same number of events for each mass point. 100k $4b$ events and 5k $6b$ events.

Zhi-Zhong's samples can be used for training.



(a)

(b)

Figure 10: Classification performance at different mass points. The AUCs of different classifiers are evaluated. Both Spa-Net and Dense-NNs are trained on the mixed datasets described in Section 26.1.

29 Two Real Singlet Model (TRSM)

The Two Real Singlet Model can be found on GitLab, which is based on the reference [7]. We consider the triple production of 125 GeV Higgs bosons via the gluon fusion:

$$gg \to h_3 \to h_2 h_1 \to h_1 h_1 h_1$$

The Higgs boson h_1 would further decay to the $b\bar{b}$ pair. We consider the banchmark point, where $m_{h_3} = 420$ GeV, $m_{h_2} = 280$ GeV, $m_{h_1} = 125$ GeV.

We use `MadGraph 3.3.1` [12] to generate these processes at a center-of-mass energy of $\sqrt{s} = 13$ TeV. The parton showering and hadronization are simulated using `Pythia 8.306` [13]. The detector simulation is conducted by `Delphes 3.4.2` [14]. Jet reconstruction is performed using `FastJet 3.3.2` [15] with the anti-k_t algorithm [16] and a jet radius of $R = 0.4$. These jets are required to have transverse momentum $p_T > 20$ GeV.

The following `MadGraph` scripts generate the TRSM Monte Carlo samples.

```
import model loop_sm_twoscalar
generate p p > iota0 > eta0 h [QCD] QCD^2<=99
output MG5/TRSM
launch MG5/TRSM

shower=Pythia8
```

```
detector=Delphes
analysis=OFF
madspin=ON
done

Cards/delphes_card.dat

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

set run_card ptb 24
set run_card etab 2.6

set spinmode none
decay eta0 > h h, (h > b b~)
decay h > b b~

done
```

Table 56 is the cutflow number at different selection cuts. These results are similar to DM-CPV's results (Table 16).

Table 56: The number of passing events, efficiencies, and passing rates for signal processes at different selection cuts.

	Count	Efficiency	Pass rate
Total	100000	1.00	1.00
≥ 6 jets	61293	0.61	0.61
≥ 4 jets with $p_{\mathrm{T}} > 40$ GeV	50282	0.82	0.50
≥ 4 b-jets	32385	0.64	0.32
Matching 3h	10040	0.31	0.10
Matching 2h	11298	0.35	0.11
Matching 1h	9017	0.28	0.09

Table 57 and 58 are the pairing performance with TRSM samples. The results are similar to the DM-CPV's (Table 17 and 18).

Table 57: χ^2 pairing efficiencies for TRSM samples. We minimize the quantity defined in Equation 3. The χ^2 method considers the possible combinations of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.239	0.493	0.000	0.362	0.145	0.614
$= 7$	0.338	0.272	0.086	0.395	0.248	0.461
≥ 8	0.423	0.155	0.109	0.389	0.347	0.357
Total	1.000	0.275	0.075	0.385	0.265	0.453

Table 58: Absolute value pairing efficiencies for TRSM samples. We minimize the quantity defined in Equation 8. The absolute value method considers the possible combination of 6 jets.

		Correctly reconstructed Higgs				
N_{Jet}	Fraction	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.239	0.438	0.000	0.378	0.184	0.564
$= 7$	0.338	0.237	0.077	0.397	0.289	0.421
≥ 8	0.423	0.141	0.100	0.390	0.370	0.337
Total	1.000	0.244	0.068	0.389	0.298	0.420

29.1 Event classifier with mixed datasets in TRSM

The $4b$ datasets consist of events from 4 benchmark points: (m_{h_3}, m_{h_2}) = (420, 280), (500, 275), (500, 300), (520, 325) GeV. The number of events for each mass point is 250k, and the total size of the $4b$ datasets is 1M.

For the jet assignment part (only for Spa-Net), the details of training and testing samples are listed as follows:

- Training sample:
  - Total sample size: 1,800,000
  - 1h sample size: 234,321
  - 2h sample size: 331,837
  - 3h sample size: 283,022
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - 1h sample size: 24,744
  - 2h sample size: 38,116
  - 3h sample size: 32,368

For event classification, both Dense-NN and Spa-Net

- Training sample:
  - Total sample size: 1,800,000
  - Signal sample size: 900,000
  - Background sample size: 900,000
  - 5% used on validation
- Testing sample:
  - Total sample size: 200,000
  - Signal sample size: 100,000
  - Background sample size: 100,000

Training SPA-NET on our server requires approximately 3.8 hours for 50 epochs.

The inputs to the SPA-NET classifiers are the four-momentum vectors of the jets in the final state. The inputs to the Dense-NN classifiers are high-level observables, which are constructed based on the jet pairing methods.

Note that, for the Dense-NN with SPA-NET pairing, we utilized the same $4b$ datasets to train both neural networks for simplicity.

Table 59 and 60 summarizes the training performance of SPA-NET using the mixed mass $4b$ tri-Higgs dataset. The training performance is worse than the one considering only one mass point.

Table 59: SPA-NET training results on the mixed mass $4b$ TRSM samples. We used the $4b$ testing sample to evaluate the pairing performance. SPA-NET is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.429	0.000	0.411	0.160	0.566
$= 7$	0.312	0.049	0.422	0.218	0.485
≥ 8	0.205	0.084	0.422	0.290	0.401
Total	0.283	0.056	0.420	0.242	0.460

Table 60: The SPA-NET classification training results with mixed mass $4b$ TRSM sample.

	Test on $4b$ datasets		Test on $6b$ datasets	
	ACC	AUC	ACC	AUC
SPA-NET	0.919	0.976	0.926	0.980

Table 61 shows the Dense-NN training results for the mixed mass $4b$ datasets. SPA-NET still outperforms Dense-NN, which is consistent with the previous results.

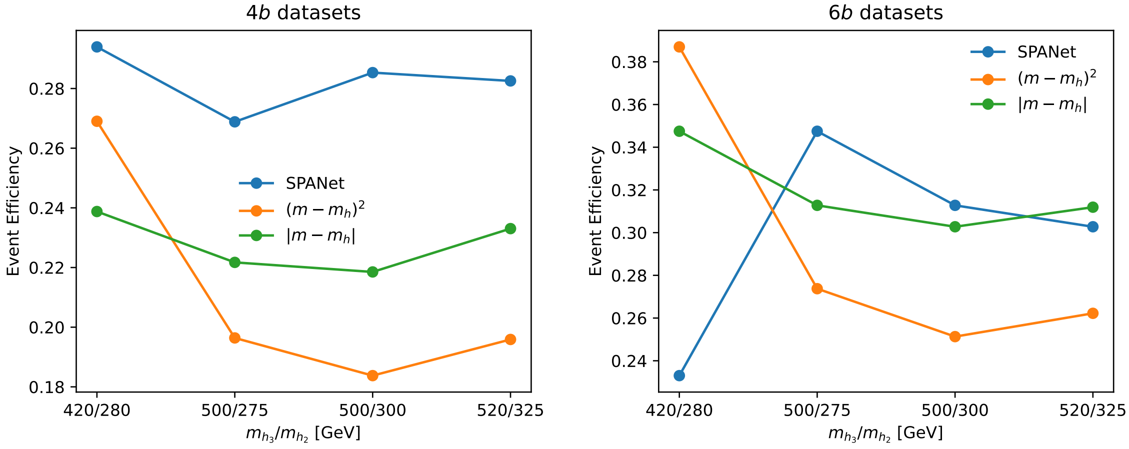
29.2 Performance at different mass points

We evaluate the pairing and classification performance at each mass point to investigate the detailed results for mixed datasets.

Figure 11 illustrates the pairing performance across different mass points. For the $4b$ dataset, SPA-NET has the highest event efficiency across all mass points. However, for the $6b$ dataset, SPA-NET does not always outperform the cut-based methods. One possible reason is that the training sample size is insufficient.

Table 61: The dense neural network training results. The ACC and AUC are evaluated based on 10 training.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.7966 ± 0.0007	0.8759 ± 0.0007	0.8351 ± 0.0004	0.9043 ± 0.0006
χ^2	0.7949 ± 0.0008	0.8740 ± 0.0007	0.8330 ± 0.0007	0.9019 ± 0.0007
Spa-Net	0.8041 ± 0.0007	0.8857 ± 0.0005	0.8329 ± 0.0006	0.9052 ± 0.0003

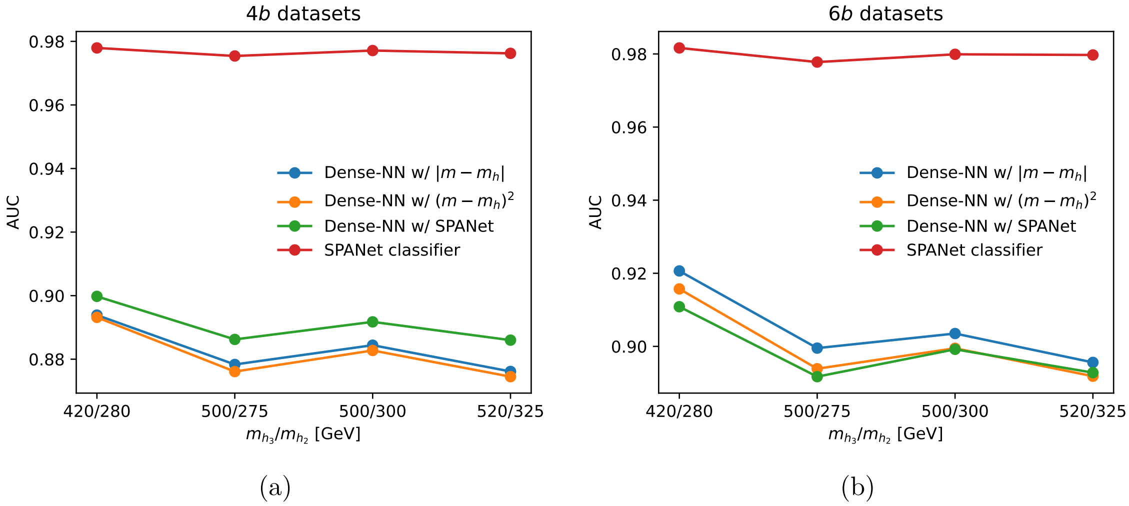


(a)

(b)

Figure 11: Pairing performance at different mass points. We evaluate the event efficiency of various pairing methods. The Spa-Net is trained on the mixed datasets described in Section 29.1.

Figure 14 shows the classification performance at different mass points. The Spa-Net classifier outperforms the Dense-NN classifier for both $4b$ and $6b$ datasets. Dense-NN classifiers with different pairing methods show similar performance trends. These results are consistent with tables 60 and 61.



(a)

(b)

Figure 12: Classification performance at different mass points. The AUCs of different classifiers are evaluated. Both Spa-Net and Dense-NNs are trained on the mixed datasets described in Section 26.1.

30 Enlarge the training samples for TRSM

We consider the same mass point mentioned in section 29.1 but increase the number of events for each mass point to 750k, and the total size of the $4b$ datasets is 3M.

For the jet assignment part (only for Spa-Net), the details of training and testing samples are listed as follows:

- Training sample:
  - Total sample size: 3,600,000
  - 1h sample size: 704,163
  - 2h sample size: 993,213
  - 3h sample size: 850,587
  - 5% used on validation

- Testing sample:
  - Total sample size: 400,000
  - 1h sample size: 74,708
  - 2h sample size: 113,360
  - 3h sample size: 96,180

For event classification, both Dense-NN and Spa-Net

- Training sample:
  - Total sample size: 3,600,000
  - Signal sample size: 2,700,000
  - Background sample size: 900,000
  - 5% used on validation
- Testing sample:
  - Total sample size: 400,000
  - Signal sample size: 300,000
  - Background sample size: 100,000

Training Spa-Net on our server requires approximately 10.1 hours for 50 epochs.

Table 62, 63 and 64 summarizes the training performance of Spa-Net using the mixed mass $4b$ tri-Higgs dataset. The pairing performance is better than table 59 by 3%. The classification results are similar to table 60. It seems the classification performance is saturated.

Table 65 shows the Dense-NN training results for the mixed mass $4b$ datasets. Spa-Net still outperforms Dense-NN, which is consistent with the previous results (table 60 and 61).

30.1 Performance at different mass points

We evaluate the pairing and classification performance at each mass point to investigate the detailed results for mixed datasets.

Figure 13 illustrates the pairing performance across different mass points. For the $4b$ dataset, Spa-Net has the highest event efficiency across all mass points. However, for the $6b$

Table 62: SPA-NET training results on the mixed mass $4b$ TRSM samples. For each mass point, we used 750k events. We used the $4b$ testing sample to evaluate the pairing performance. SPA-NET is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.445	0.000	0.388	0.167	0.574
$= 7$	0.337	0.054	0.397	0.211	0.506
≥ 8	0.238	0.088	0.404	0.271	0.431
Total	0.311	0.060	0.398	0.231	0.483

Table 63: SPA-NET training results on the mixed mass $4b$ TRSM samples. For each mass point, we used 750k events. We used the $6b$ testing sample to evaluate the pairing performance. SPA-NET is trained on jet pairing and event classification tasks at the same time.

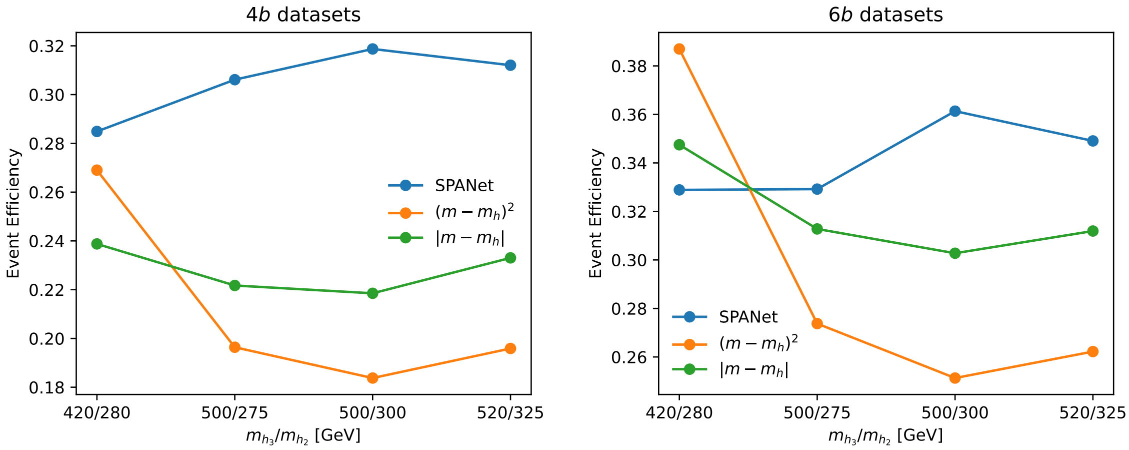
	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.479	0.000	0.372	0.149	0.603
$= 7$	0.387	0.033	0.390	0.190	0.539
≥ 8	0.267	0.066	0.400	0.267	0.444
Total	0.343	0.044	0.392	0.221	0.503

Table 64: The SPA-NET classification training results with mixed mass $4b$ TRSM sample.

	Test on $4b$ datasets		Test on $6b$ datasets	
	ACC	AUC	ACC	AUC
SPA-NET	0.932	0.974	0.921	0.977

Table 65: The dense neural network training results. The ACC and AUC are evaluated based on 10 training.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.8597 ± 0.0001	0.8763 ± 0.0006	0.8345 ± 0.0005	0.9047 ± 0.0005
χ^2	0.8585 ± 0.0001	0.8740 ± 0.0004	0.8330 ± 0.0006	0.9022 ± 0.0006
SPA-NET	0.8640 ± 0.0002	0.8881 ± 0.0005	0.8361 ± 0.0004	0.9082 ± 0.0003



(a)

(b)

Figure 13: Pairing performance at different mass points. We evaluate the event efficiency of various pairing methods. The SPA-NET is trained on the mixed datasets described in Section 30.

dataset, SPA-NET does not always outperform the cut-based methods. One possible reason is that the training sample size is insufficient.

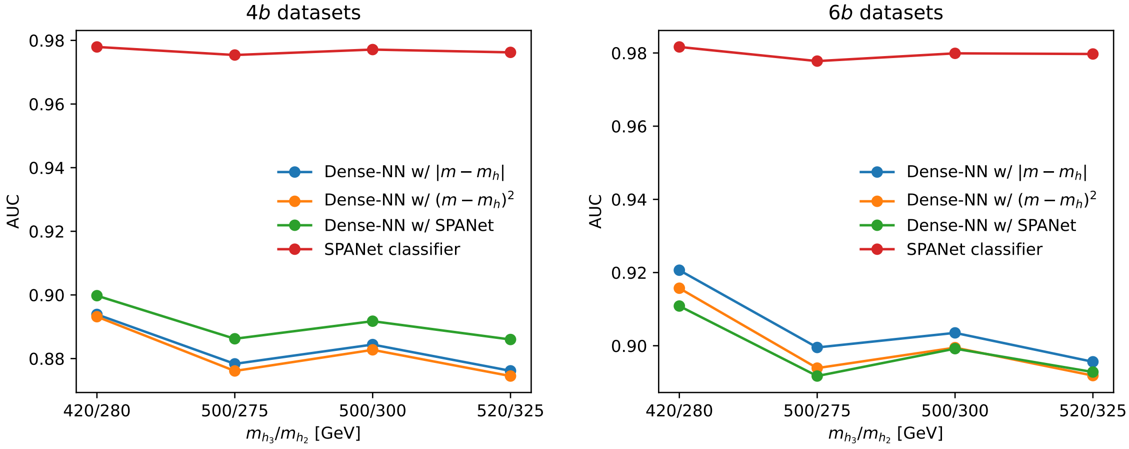
Figure 14 shows the classification performance at different mass points. The SPA-NET classifier outperforms the Dense-NN classifier for both 4b and 6b datasets. Dense-NN classifiers with different pairing methods show similar performance trends. These results are consistent with tables 64 and 65.

31 TRSM: 5 mass points

This section considers the 5 mass points: (m_{h_3}, m_{h_2}) = (420, 280), (500, 275), (500, 300), (520, 325), (500, 350) GeV. In the training set, each signal point contains 250k events, and the background includes 1M events. Thus, the total size of the 4b training datasets is 2.25M.

For the jet assignment part (only for SPA-NET), the details of training and testing samples are listed as follows:

- Training sample:
  - Total sample size: 2,250,000
  - 1h sample size: 299,919



(a) (b)

Figure 14: Classification performance at different mass points. The AUCs of different classifiers are evaluated. Both Spa-Net and Dense-NNs are trained on the mixed datasets described in Section 26.1.

  - 2h sample size: 496,944
  - 3h sample size: 394,491
  - 5% used on validation

- Testing sample:
  - Total sample size: 300,000
  - 1h sample size: 39,490
  - 2h sample size: 129,085
  - 3h sample size: 79,385

For event classification, both Dense-NN and Spa-Net

- Training sample:
  - Total sample size: 2,250,000
  - Signal sample size: 1,250,000
  - Background sample size: 1,000,000
  - 5% used on validation
- Testing sample:

- Total sample size: 300,000
- Signal sample size: 250,000
- Background sample size: 50,000

Training SPA-NET on our server requires approximately 6.25 hours for 50 epochs.

Table 66, 67 and 68 summarizes the training performance of SPA-NET using the mixed mass $4b$ dataset. The pairing performance is worse than table 62 and 63. The classification results are similar to table 64. It seems the classification performance is saturated.

Table 66: SPA-NET training results on the mixed mass $4b$ TRSM samples. For each mass point, we used 250k events. We used the $4b$ testing sample to evaluate the pairing performance. SPA-NET is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.343	0.000	0.451	0.206	0.494
$= 7$	0.282	0.058	0.415	0.245	0.459
≥ 8	0.175	0.082	0.421	0.322	0.370
Total	0.244	0.057	0.426	0.273	0.424

Table 67: SPA-NET training results on the mixed mass $4b$ TRSM samples. For each mass point, we used 250k events. We used the $6b$ testing sample to evaluate the pairing performance. SPA-NET is trained on jet pairing and event classification tasks at the same time.

	Correctly reconstructed Higgs				
N_{Jet}	3h	2h	1h	0h	Higgs Efficiency
$= 6$	0.399	0.000	0.416	0.185	0.538
$= 7$	0.305	0.029	0.423	0.243	0.465
≥ 8	0.212	0.057	0.417	0.314	0.389
Total	0.274	0.038	0.418	0.269	0.439

Table 68: The SPA-NET classification training results with mixed mass $4b$ TRSM sample.

	Test on $4b$ datasets		Test on $6b$ datasets	
	ACC	AUC	ACC	AUC
SPA-NET	0.946	0.974	0.920	0.977

Table 69 shows the Dense-NN training results for the mixed mass $4b$ datasets. SPA-NET still outperforms Dense-NN, which is similar to the previous exercise (table 64 and 65).

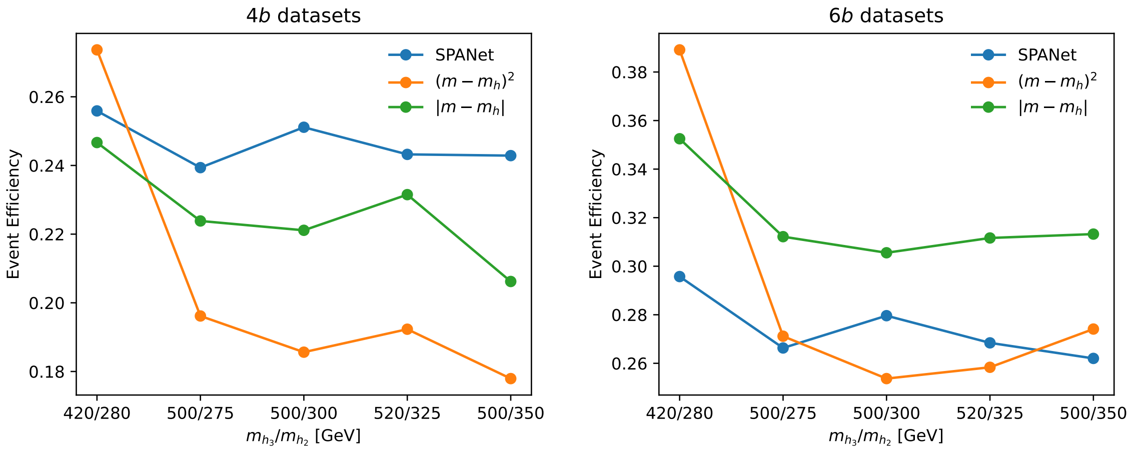
Table 69: The dense neural network training results. The ACC and AUC are evaluated based on 10 training.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.8936 ± 0.0002	0.8807 ± 0.0002	0.8329 ± 0.0006	0.9023 ± 0.0005
χ^2	0.8933 ± 0.0002	0.8786 ± 0.0003	0.8308 ± 0.0008	0.9001 ± 0.0006
SPA-NET	0.8948 ± 0.0001	0.8843 ± 0.0002	0.8308 ± 0.0005	0.9029 ± 0.0003

31.1 Performance at different mass points

We evaluate the pairing and classification performance at each mass point to investigate the detailed results for mixed datasets.

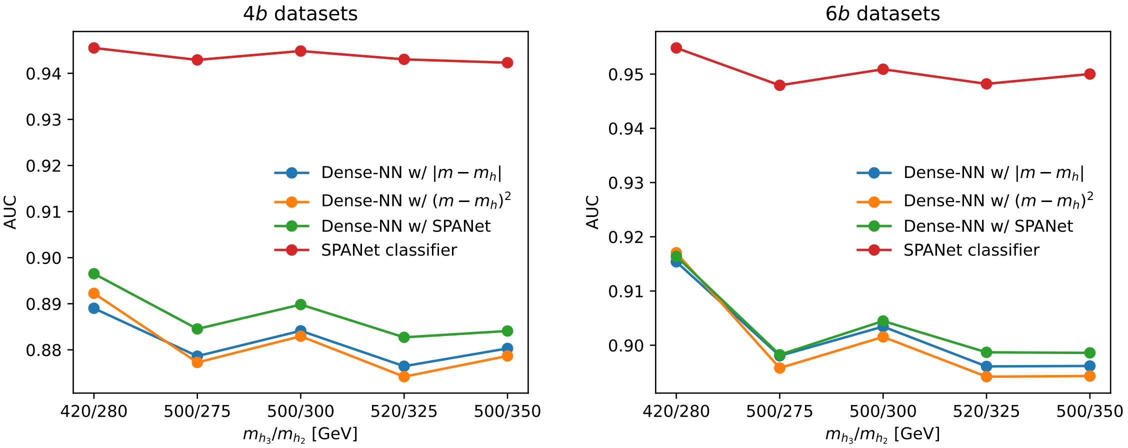
Figure 15 illustrates the pairing performance across different mass points. For the $4b$ dataset, SPA-NET has the highest event efficiency except the $(420, 280)$ GeV point. However, for the $6b$ dataset, SPA-NET does not perform better than the cut-based methods. One possible reason is that the training sample size is insufficient.



(a) (b)

Figure 15: Pairing performance at different mass points. We evaluate the event efficiency of various pairing methods. The SPA-NET is trained on the mixed datasets described in section 31.

Figure 16 shows the classification performance at different mass points. The Spa-Net classifier outperforms the Dense-NN classifier for both $4b$ and $6b$ datasets. Dense-NN classifiers with different pairing methods show similar performance trends. These results are consistent with tables 68 and 69.



(a)

(b)

Figure 16: Classification performance at different mass points. The AUCs of different classifiers are evaluated. Both Spa-Net and Dense-NNs are trained on the mixed datasets described in Section 31.

31.2 Upper limit of cross-section

The $\mathrm{CL_s}$ method is used to set the upper limits of the cross-section. The signal strength μ_s is chosen as the parameter of interest (POI). The POI is excluded at the 95% confidence level when the $\mathrm{CL_s} < 0.05$. The package `pyhf` [10, 11] is utilized to calculate the upper limit. Once the upper limit of signal strength is obtained, it can be converted to the upper limit of the cross-section.

Figure 17 shows the 95% CL upper limits for cross-section $\sigma\,(pp \to hhh)$. Dense-NN classifiers with different pairing methods demonstrate similar results. The Spa-Net classifier performs the best among all selection methods. Note that these values are preliminary, as we need to use the correct background cross-sections to normalize them. The final results will differ by an overall factor.

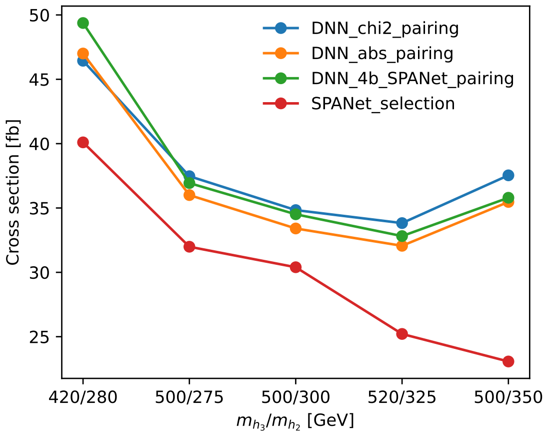


Figure 17: 95% CL upper limits of the cross-section $\sigma\,(pp \rightarrow hhh)$.

32 TRSM: 5 mass points, 1M datasets

This section considers the same mass points as section 31 and increases the training sample size. In the training set, each signal point contains 1M events, and the background includes 1M events. Thus, the total size of the $4b$ training datasets is 6M.

For the jet assignment part (only for Spa-Net), the details of training and testing samples are listed as follows:

- Training sample:
  - Total sample size: 6,000,000
  - 1h sample size: 1,200,697
  - 2h sample size: 1,985,367
  - 3h sample size: 1,579,269
  - 5% used on validation
- Testing sample:
  - Total sample size: 300,000
  - 1h sample size: 59,966
  - 2h sample size: 99,105
  - 3h sample size: 79,240

For event classification, both Dense-NN and Spa-Net

- Training sample:

  - Total sample size: 6,000,000
  - Signal sample size: 5,000,000
  - Background sample size: 1,000,000
  - 5% used on validation

- Testing sample:

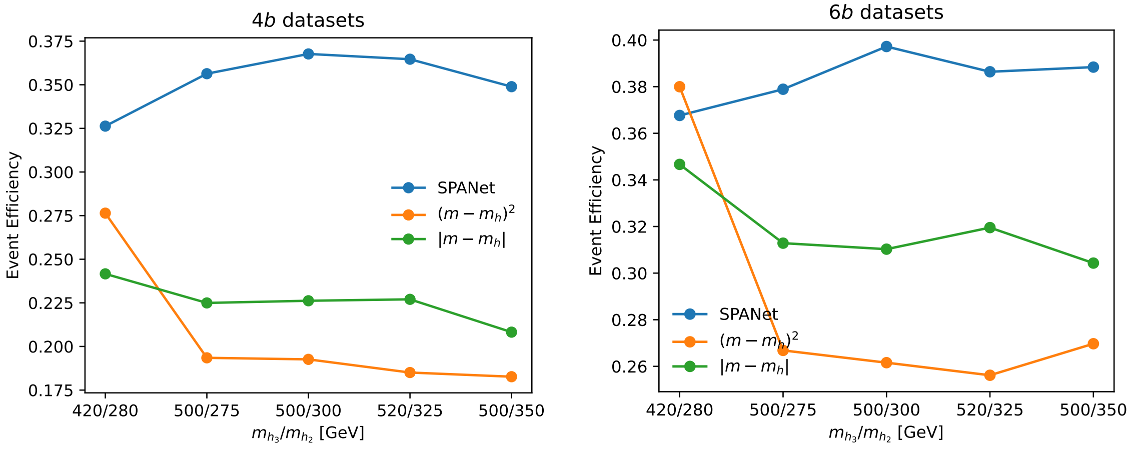
  - Total sample size: 300,000
  - Signal sample size: 250,000
  - Background sample size: 50,000

Training SPA-NET on our server requires approximately 16.04 hours for 50 epochs.

32.1 Performance at different mass points

We evaluate the pairing and classification performance at each mass point to investigate the detailed results for mixed datasets.

Figure 18 illustrates the pairing performance across different mass points. For the $4b$ dataset, SPA-NET has the highest event efficiency. For the $6b$ dataset, SPA-NET still performs better than the cut-based methods except at the point (420, 280) GeV.

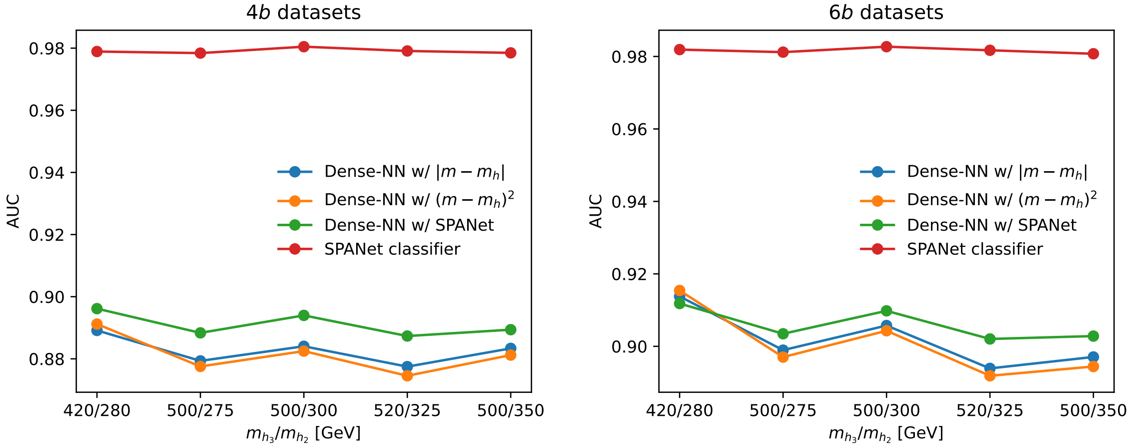


(a)

(b)

Figure 18: Pairing performance at different mass points. We evaluate the event efficiency of various pairing methods. The SPA-NET is trained on the mixed datasets described in section 32.

Figure 19 shows the classification performance at different mass points. The SPA-NET classifier outperforms the Dense-NN classifier for both $4b$ and $6b$ datasets. Dense-NN classifiers with different pairing methods show similar performance trends. These results are similar to the previous ones.



(a)

(b)

Figure 19: Classification performance at different mass points. The AUCs of different classifiers are evaluated. Both SPA-NET and Dense-NNs are trained on the mixed datasets described in Section 31.

32.2 Upper limit of cross-section

The CL$_s$ method is used to set the upper limits of the cross-section. The signal strength μ_s is chosen as the parameter of interest (POI). The POI is excluded at the 95% confidence level when the CL$_s$ < 0.05. The package `pyhf` [10, 11] is utilized to calculate the upper limit. Once the upper limit of signal strength is obtained, it can be converted to the upper limit of the cross-section.

Figure 20 shows the 95% CL upper limits for cross-section $\sigma\,(pp \to hhh)$. Dense-NN classifiers with different pairing methods demonstrate similar results. The SPA-NET classifier performs the best among all selection methods. Note that these values are preliminary, as we need to use the correct background cross-sections to normalize them. The final results will differ by an overall factor.

The results of the SPA-NET classifier are nearly identical to those shown in figure 17, as the upper limit significantly depends on the classification results, and the classification performance is consistent across both the 250k and 1M datasets.

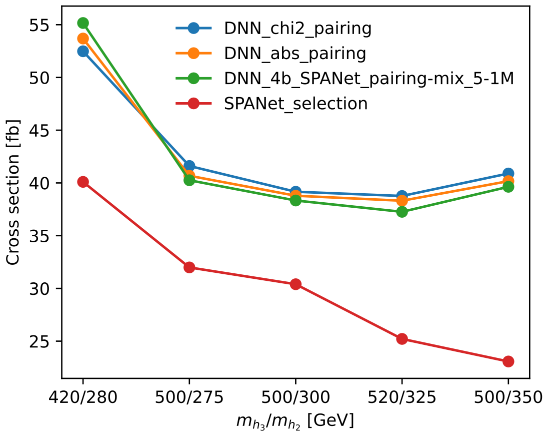


Figure 20: 95% CL upper limits of the cross-section $\sigma\,(pp \to hhh)$.

32.3 Update the upper limit

The parameter setting:

- Luminosity: $\mathcal{L} = 126$ fb^{-1}.
- Binning: Divide $[0, 1]$ into 20 uniform bins.

Figure 21 shows the 95% CL upper limits for cross-section $\sigma\,(pp \to hhh)$. Here, we used the correct background cross-section. We also include the ATLAS values for comparison.

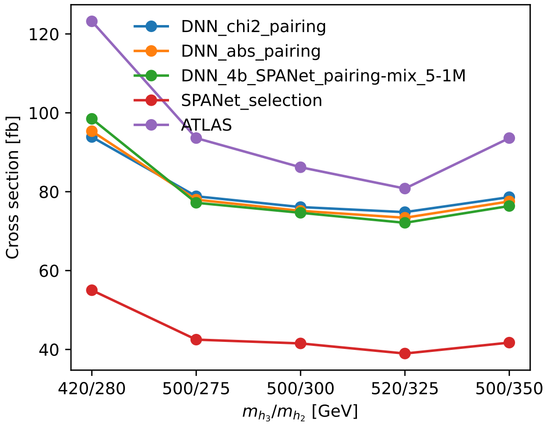


Figure 21: 95% CL upper limits of the cross-section $\sigma\,(pp \to hhh)$.

32.4 Modified the Dense-NN training setting

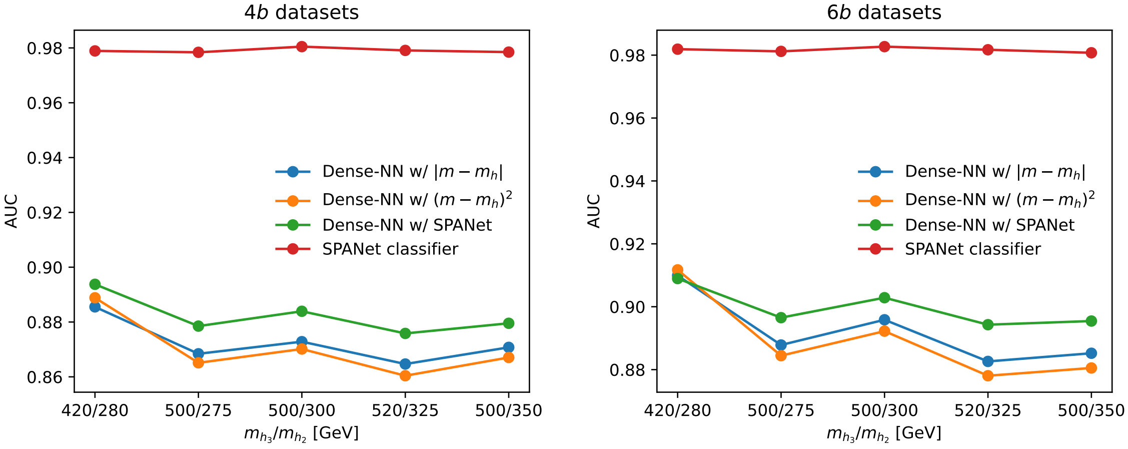
To make a fari comparison with ATLAS results, we change the number of hidden nodes to 24 for each layer. We reduced the training sample size to 900k, such the number of background event is closed to 5b data. The class weights are also added.

Table 70 shows the Dense-NN training results for the mixed mass $4b$ datasets. The training results are worse than 1M datasets.

Table 70: The dense neural network training results. The ACC and AUC are evaluated based on 10 training.

	Test on $4b$ datasets		Test on $6b$ datasets	
Pairing method	ACC	AUC	ACC	AUC
Absolute value	0.8927 ± 0.0002	0.8724 ± 0.0006	0.8255 ± 0.0008	0.8929 ± 0.0009
χ^2	0.8924 ± 0.0005	0.8707 ± 0.0017	0.8241 ± 0.0013	0.8898 ± 0.0014
SPA-NET	0.8956 ± 0.0002	0.8821 ± 0.0008	0.8275 ± 0.0007	0.8990 ± 0.0008

Figure 22 shows the classification performance at different mass points. Dense-NN classifiers with different pairing methods show similar performance trends. These results are worse than figure 19 by 1%.



(a)

(b)

Figure 22: Classification performance at different mass points. The AUCs of different classifiers are evaluated. Both SPA-NET and Dense-NNs are trained on the mixed datasets described in Section 31.

Figure 23 shows the 95% CL upper limits for cross-section $\sigma\,(pp \to hhh)$. The upper limits of Dense-NN are similar to figure 21.

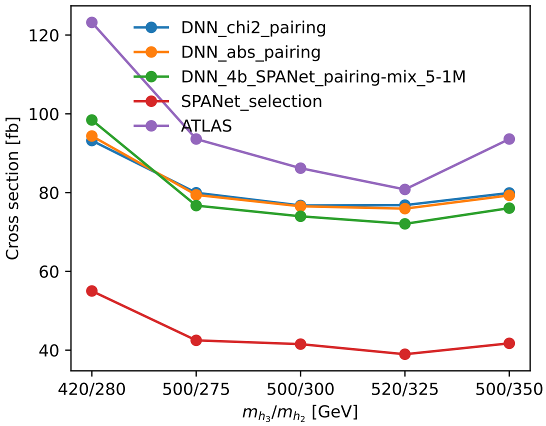


Figure 23: 95% CL upper limits of the cross-section $\sigma\,(pp \to hhh)$.

References

[1] T.-K. Chen, C.-W. Chiang, and I. Low, “Simple model of dark matter and CP violation,” *Phys. Rev. D*, vol. 105, no. 7, p. 075025, 2022.

[2] M. J. Fenton, A. Shmakov, T.-W. Ho, S.-C. Hsu, D. Whiteson, and P. Baldi, “Permutationless many-jet event reconstruction with symmetry preserving attention networks,” *Phys. Rev. D*, vol. 105, p. 112008, Jun 2022.

[3] M. J. Fenton, A. Shmakov, H. Okawa, Y. Li, K.-Y. Hsiao, S.-C. Hsu, D. Whiteson, and P. Baldi, “Extended Symmetry Preserving Attention Networks for LHC Analysis,” 9 2023.

[4] A. Shmakov, M. J. Fenton, T.-W. Ho, S.-C. Hsu, D. Whiteson, and P. Baldi, “SPANet: Generalized permutationless set assignment for particle physics using symmetry preserving attention,” *SciPost Phys.*, vol. 12, p. 178, 2022.

[5] A. Papaefstathiou, G. Tetlalmatzi-Xolocotzi, and M. Zaro, “Triple Higgs boson production to six b-jets at a 100 TeV proton collider,” *Eur. Phys. J. C*, vol. 79, no. 11, p. 947, 2019.

[6] A. Papaefstathiou and G. Tetlalmatzi-Xolocotzi, “Multi-Higgs Boson Production with Anomalous Interactions at Current and Future Proton Colliders,” 12 2023.

[7] A. Papaefstathiou, T. Robens, and G. Tetlalmatzi-Xolocotzi, “Triple Higgs Boson Production at the Large Hadron Collider with Two Real Singlet Scalars,” *JHEP*, vol. 05, p. 193, 2021.

[8] B. Stanislaus, *Searching for Beyond the Standard Model resonances in the $HH \to b\bar{b}b\bar{b}$ final state using the ATLAS detector.* PhD thesis, Oxford U., 2020.

[9] R. D. Ball *et al.*, "Parton distributions for the LHC Run II," *JHEP*, vol. 04, p. 040, 2015.

[10] L. Heinrich, M. Feickert, and G. Stark, "pyhf: v0.7.3." https://github.com/scikit-hep/pyhf/releases/tag/v0.7.3.

[11] L. Heinrich, M. Feickert, G. Stark, and K. Cranmer, "pyhf: pure-python implementation of histfactory statistical models," *Journal of Open Source Software*, vol. 6, no. 58, p. 2823, 2021.

[12] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *JHEP*, vol. 07, p. 079, 2014.

[13] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An introduction to PYTHIA 8.2," *Comput. Phys. Commun.*, vol. 191, pp. 159–177, 2015.

[14] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment," *JHEP*, vol. 02, p. 057, 2014.

[15] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur. Phys. J. C*, vol. 72, p. 1896, 2012.

[16] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *JHEP*, vol. 04, p. 063, 2008.