Eur. Phys. J. C (2018) 78:1010
https://doi.org/10.1140/epjc/s10052-018-6455-z

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Signal mixture estimation for degenerate heavy Higgses using a deep neural network

Anders Kvellestad[1,2,a], **Steffen Maeland**[3,b], **Inga Strümke**[3,c]

[1] Department of Physics, University of Oslo, 0316 Oslo, Norway
[2] Blackett Laboratory, Department of Physics, Imperial College London, Prince Consort Road, London SW7 2AZ, UK
[3] Department of Physics and Technology, University of Bergen, 5020 Bergen, Norway

Received: 9 May 2018 / Accepted: 15 November 2018 / Published online: 12 December 2018

Abstract If a new signal is established in future LHC data, a next question will be to determine the signal composition, in particular whether the signal is due to multiple near-degenerate states. We investigate the performance of a deep learning approach to signal mixture estimation for the challenging scenario of a ditau signal coming from a pair of degenerate Higgs bosons of opposite CP charge. This constitutes a parameter estimation problem for a mixture model with highly overlapping features. We use an unbinned maximum likelihood fit to a neural network output, and compare the results to mixture estimation via a fit to a single kinematic variable. For our benchmark scenarios we find a $\sim 20\%$ improvement in the estimate uncertainty.

1 Introduction

Machine learning techniques have already proven useful in particle physics, especially for separating signal from background events in analyses of LHC data. More recently, *deep learning* methods, such as multi-layer neural networks, have been shown to perform very well, due to their ability to learn complex non-linear correlations in high-dimensional data [1–3]. In this paper we study the performance of a deep neural network classifier, but rather than classifying signal vs. background we focus on estimating the mixture of different signal classes in a dataset. This is motivated by the not-unlikely scenario where a new (and possibly broad) resonance is discovered in future LHC data, but limited statistics makes the interpretation difficult, in particular the question of whether the signal is due to multiple degenerate states. In such a scenario it will clearly be important to squeeze as much information as possible from the available data.

While the approach studied here is general, we take a Two-Higgs-Doublet Model (THDM) as our example scenario. In these models the Higgs sector of the Standard Model (SM) is extended with an additional $SU(2)$ doublet, predicting the existence of a pair of charged scalars ($H^{\pm}$) and three neutral scalars (h, H, A), one of which should be the observed 125 GeV Higgs. Several more extensive frameworks for New Physics predict a Higgs sector with the structure of a THDM, the prime example being the Minimal Supersymmetric Standard Model (MSSM). A further motivation for THDMs comes from the fact that the extended scalar sector can allow for additional sources of CP violation and a strongly first-order electroweak phase transition, as required for electroweak baryogenesis [4–7]. For a recent study of this, see [8].

We associate the light scalar h with the observed 125 GeV Higgs and take the heavier scalars H, A and $H^{\pm}$ to be mass degenerate. The focus of our study is on the ditau LHC signal from decays of the neutral states H and A, which in this case are indistinguishable save for their opposite CP charges. Searches for heavy neutral Higgses in ditau final states are carried out by both the ATLAS and CMS collaborations, see [9,10] for recent results.

The remainder of this paper is structured as follows. In Sect. 2 we motivate why it is reasonable to expect a certain level of mass-degeneracy among the new scalars in THDMs and present our example THDM scenario. The technical setup for our analysis is given in Sect. 3. Here we define our signal models, describe the procedure for Monte Carlo event generation and detail the neural network layout and training. In Sect. 4 we demonstrate H/A signal mixture estimation using the method of fitting a single kinematic variable. The result serves as our baseline for judging the performance of the deep learning approach. Our main results are presented in

[a] e-mail: anders.kvellestad@fys.uio.no
[b] e-mail: steffen.maeland@uib.no
[c] e-mail: inga.strumke@uib.no

Sect. 5. Here we estimate the signal mixture via a maximum likelihood fit to the output distribution from a network trained to separate H and A ditau events. The results are compared to those from Sect. 4. We state our conclusions in Sect. 6.

2 Theory and motivation

The starting point for our study is a THDM scenario where $m_H \approx m_A$. Our main motivation for this choice is to obtain a challenging test case for signal mixture estimation. However, there are also physical reasons to expect the H and A states to have similar masses. After requiring that the scalar potential has a minimum in accordance with electroweak symmetry breaking, we are left with a model with only two mass scales, $v \approx 246\,\text{GeV}$ and a free mass parameter μ, to control the four masses m_h, m_H, m_A and $m_{H^\pm}$. From the point of view of the general THDM parameter space, the least fine-tuned way to align the light state h with SM predictions, as favoured by LHC Higgs data, is to move towards simultaneous decoupling of the three heavier states by increasing μ, leaving v to set the scale for $m_h = 125\,\text{GeV}$ [11]. This points to a scenario where $|m_H - m_A| \lesssim 100\,\text{GeV}$, and quite possibly much smaller, depending on the quartic couplings of the scalar potential.[1]

Further motivation for a small H–A mass difference can be found in less general realisations of THDMs. For the type-II THDM in the MSSM the quartic couplings are fixed by the squares of the SM gauge couplings, resulting in the tree-level prediction that $m_H - m_A \lesssim 10\,\text{GeV}$ for $m_A \sim 400\,\text{GeV}$ and $\tan\beta \sim 1$, and decreasing further with increasing $\tan\beta$ or m_A [13]. Another well-motivated scenario predicting closely degenerate H and A states is the $SO(5)$-based Maximally Symmetric THDM [14].

When mass degenerate, the H and A appear identical except for their CP charge. If the properties of the light h deviates from SM predictions, this difference in CP charge can manifest as non-zero ZZ and WW couplings for H, while for the CP-odd A the Zh coupling is available. However, these couplings all vanish in the perfect SM-alignment limit we assume here. Yet the CP nature of H and A is still expressed as spin correlations in fermionic decay modes, impacting the kinematics of subsequent decays. Here we study the channels $H \to \tau\tau$ and $A \to \tau\tau$. Methods for reconstructing spin correlations in ditau decays of the 125 GeV Higgs have been investigated in detail [15–18], providing a good baseline for comparison. The use of neural networks to optimize CP measurements for the 125 GeV state is studied in [19,20].

2.1 Benchmark scenario

Two-Higgs-Doublet Models are classified in different types based on the structure of the Yukawa sector. We choose a benchmark scenario within the CP-conserving lepton-specific THDM, with $m_H = m_A = m_{H^\pm} = 450\,\text{GeV}$. In this model, the quarks couple to one of the Higgs doublets and the leptons to the other. This enables large branching ratios for $H/A \to \tau\tau$, even for masses above the 350 GeV threshold for $H/A \to t\bar{t}$.

By varying the remaining THDM parameters we can obtain a wide range of ditau signal strengths for the H and A states at 450 GeV. In Appendix A we illustrate how $\sigma(pp \to A) \times \mathcal{B}(A \to \tau\tau)$ and $\sigma(pp \to H) \times \mathcal{B}(H \to \tau\tau)$ vary across the high-mass region of the lepton-specific THDM parameter space. For $m_H = m_A \approx 450\,\text{GeV}$, we find that the ditau signal strengths can reach up to $\sigma(pp \to H) \times \mathcal{B}(H \to \tau\tau) \approx 34\,\text{fb}$ and $\sigma(pp \to A) \times \mathcal{B}(A \to \tau\tau) \approx 54\,\text{fb}$ in 13 TeV proton–proton collisions. This includes production via gluon–gluon fusion and bottom-quark annihilation, with cross sections evaluated at NLO using SusHi 1.6.1 [21–27] and branching ratios obtained from 2HDMC 1.7.0 [28].

For comparison, in Appendix A we also show the result of a similar scan of the type-I THDM. In this model all fermions couple to only one of the two Higgs doublets. Compared to the lepton-specific THDM, the ditau signal in type-I THDM suffers a much stronger suppression from the $H/A \to t\bar{t}$ channel.

As further described in Sects. 3 and 4, the mixture estimation techniques we study require each tau to decay through the $\tau^\pm \to \pi^\pm\pi^0\nu$ channel, which has a branching ratio of 25%. However, the neural network method we employ can be extended to include other tau decay modes as well, by implementing the "impact parameter method" in [18] in addition to the "ρ decay-plane method" used here.

If we only assume the $\tau^\pm \to \pi^\pm\pi^0\nu$ decay channel and an acceptance times efficiency of 5%–10% for the signal selection, our example scenarios predict no more than ~ 100 signal events for the anticipated 300 fb^{-1} dataset at the end of Run 3. However, as the model scan in Appendix A shows, considering slightly lower benchmark masses can provide an order of magnitude increase in the predicted cross-section. Also, extending the method to include more tau decay channels can greatly increase the statistics available to the analysis discussed here. Still, the large backgrounds in the ditau channel, e.g. from "fake QCD taus", implies that a signal mixture estimation study for the THDM benchmark scenario we present here likely will require the improved statistics of the full High-Luminosity LHC dataset.

[1] A large H–A mass difference in this decoupling scenario relies on $\mathcal{O}(1)$ quartic couplings. We note that when loop corrections are taken into account, the viability of such scenarios can be significantly more restricted than what tree-level results suggest [12].

We do not include a third mixture component representing ditau backgrounds for our benchmark study. Clearly, the inclusion of backgrounds will increase the uncertainty in the estimated $H/A \to \tau\tau$ signal mixture. However, as we discuss in more detail in Sect. 5.1, the mixture estimate obtained from the neural network approach we study here is likely to be less affected by backgrounds than traditional mixture estimation from fitting a single kinematic variable.

For our further discussions we define the parameter α as the ratio of the $A \to \tau\tau$ signal strength to the total ditau signal strength,

$$\alpha \equiv \frac{\sigma(pp \to A) \times \mathcal{B}(A \to \tau\tau)}{\sigma(pp \to A) \times \mathcal{B}(A \to \tau\tau) + \sigma(pp \to H) \times \mathcal{B}(H \to \tau\tau)}. \tag{1}$$

This is the parameter we seek to determine in our signal mixture estimation.[2] The parameter region of our benchmark scenario predicts values of α between 0.5 and 0.7. To allow for some further variation in the assumptions, we will in our tests use α values of 0.5, 0.7 and 0.9.

3 Analysis setup

3.1 Event generation

We generate 13 TeV pp Monte Carlo events for this study using Pythia 8.219 [29,30]. Only gluon-gluon fusion and bottom-quark annihilation are considered, as these are the dominant H/A production modes at the LHC.[3] For our analysis we select opposite-sign taus decaying to $\pi^{\pm}\pi^0\nu$, which is the decay mode with the highest branching ratio. In order to roughly match recent LHC searches for $H/A \to \tau\tau$, taus are required to have visible transverse momentum p_{T} larger than 40 GeV and pseudorapidity less than 2.1. Further, we require the taus to be separated by $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} > 0.5$, and that there are no more than two taus in the event which pass the p_{T} selection. Events with muons or electrons with $p_{\mathrm{T}} > 20$ GeV are rejected.

Detector effects are taken into account by randomly smearing the directions and energies of the outgoing pions, following the procedure described in [18]: Each track is deflected by a random polar angle θ, which is drawn from a Gaussian distribution with width σ_θ, so that the smeared track lies within a cone around the true track direction. For charged pions a value of $\sigma_\theta = 1$ mrad is used, while the energy resolution is $\Delta E/E = 5\%$. For neutral pions, we use $\sigma_\theta = 0.025/\sqrt{12}$ rad and $\Delta E/E = 10\%$. To gauge the impact of such detector effects on our results, we repeat the main analyses in Sects. 4 and 5 for simulated data with and without detector smearing.

3.2 Network input features

For the neural signal mixture estimation in Sect. 5, we train a network to separate $H \to \tau\tau$ events from $A \to \tau\tau$ events. The four-momenta of the visible tau decay products ($\pi^{\pm}$ and π^0) constitute the most basic kinematic input features to our network. The momenta are boosted back to the visible ditau rest frame (the zero-momentum frame for the four pions) and rotated so that the visible taus are back-to-back along the z-axis. The system is then rotated a second time, now around the z-axis, so that the x-component of the π^+ is zero. This is done in order to align all events to a common orientation, as the azimuthal angle around the z-axis carries no physics information.

In addition to the pion momenta, the network is trained on missing transverse energy (E_T^{miss}); the invariant mass of the four-pion system (m_{vis}); the transverse mass (m_T^{tot}); the impact parameter vectors of the charged pions, which help constrain the neutrino directions; the pion energy ratios $\Upsilon^{\pm}$, defined as

$$\Upsilon^{\pm} = \frac{E_{\pi^{\pm}} - E_{\pi^0}}{E_{\pi^{\pm}} + E_{\pi^0}}; \tag{2}$$

and the angle φ^* between the tau decay planes. For φ^* we follow the definition in [18],[4] which uses the direction $\hat{\mathbf{p}}_{\perp}^{(0)\pm}$ of the π^0 transverse to the direction $\hat{\mathbf{p}}^{\pm}$ of the corresponding $\pi^{\pm}$, to form an intermediate observable $\varphi \in [0, \pi)$ and a CP-odd triple correlation product $\mathcal{O}^*$,

$$\varphi = \arccos\left(\hat{\mathbf{p}}_{\perp}^{(0)+} \cdot \hat{\mathbf{p}}_{\perp}^{(0)-}\right) \quad \text{and} \tag{3}$$

$$\mathcal{O}^* = \hat{\mathbf{p}}^{+} \cdot \left(\hat{\mathbf{p}}_{\perp}^{(0)+} \times \hat{\mathbf{p}}_{\perp}^{(0)-}\right). \tag{4}$$

From these, we can define an angle continuous on the interval $[0, 2\pi)$:

$$\varphi' = \begin{cases} \varphi & \text{if } \mathcal{O}^* \geq 0 \\ 2\pi - \varphi & \text{if } \mathcal{O}^* < 0 \end{cases}. \tag{5}$$

[2] In linking this theory quantity directly with the H/A event mixture in the datasets we simulate, we make the approximation that the acceptance times efficiency is equal for $H \to \tau\tau$ and $A \to \tau\tau$ events.

[3] The magnitudes of the up-type and down-type Yukawa couplings have the same $\tan\beta$ dependence in both the lepton-specific and the type-I THDM. Gluon–gluon fusion through a top loop is therefore by far the most important production channel for the scenarios considered here.

[4] Our definition of φ^* only differs from that in [18] in that we define φ^* in the $\pi^+\pi^0\pi^-\pi^0$ zero-momentum frame, whereas the $\pi^+\pi^-$ zero-momentum frame is used in [18].

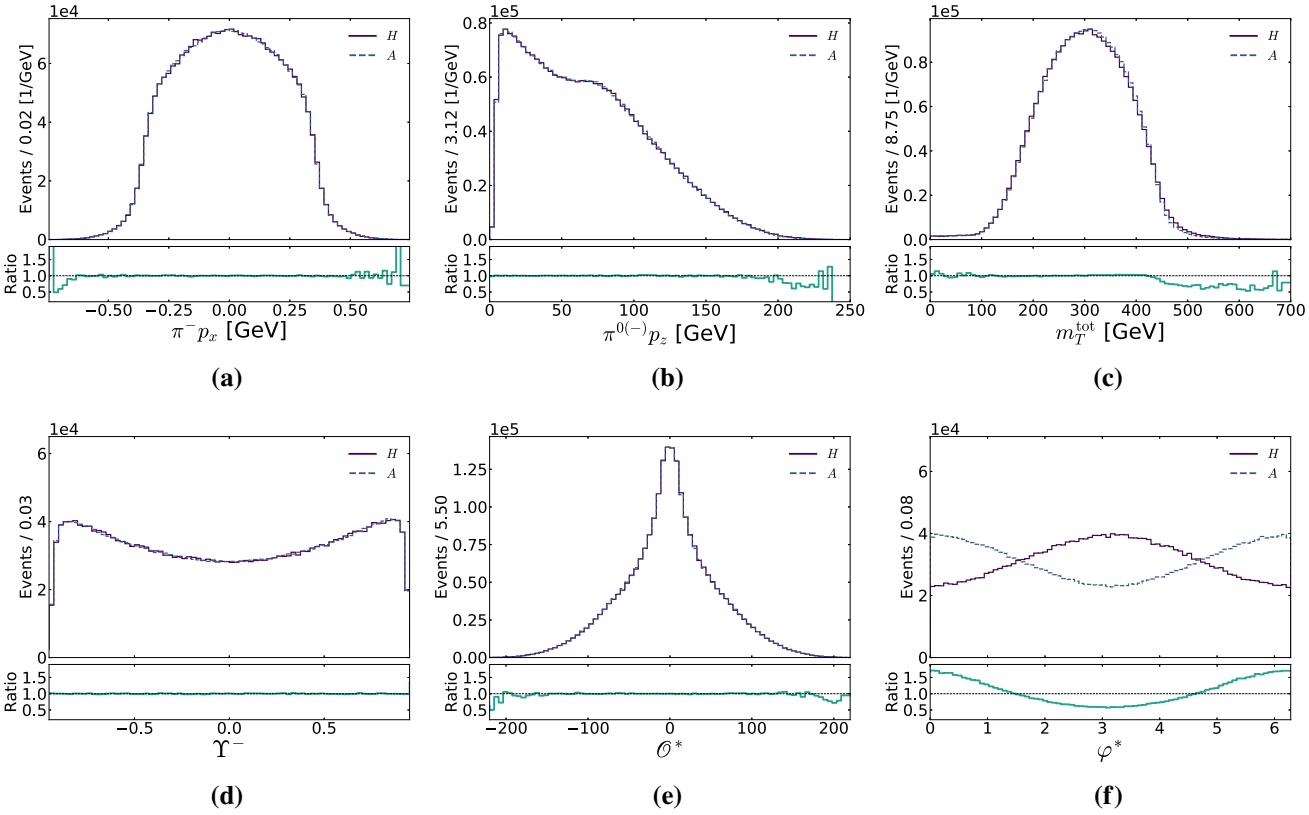


Fig. 1 Distributions for some kinematic features in $H \to \tau\tau \to (\pi^+\pi^0\nu)(\pi^-\pi^0\nu)$ events (solid purple line) and $A \to \tau\tau \to (\pi^+\pi^0\nu)(\pi^-\pi^0\nu)$ events (dashed blue line), assuming $m_H = m_A = 450\,\mathrm{GeV}$. The quantities in **a** and **b** are momentum components of the π^- and π^0 from the τ^- decay, after each event has been boosted back to the visible ditau restframe and rotated such that the taus are back-to-back in the z direction and the x-component of the π^+ momentum is zero. **c** The transverse mass m_T^{tot}, defined in [31]. The observables Υ (**d**) and $\mathcal{O}^*$ (**e**), defined in Eqs. (2) and (4), respectively, are required for the computation of φ^*, along with the momentum vectors of the tau decay products. The distribution of φ^* is shown in (**f**). The green graph below each plot shows the ratio of the A-event and H-event distributions

The distribution of φ' depends on the sign of the product $\Upsilon^+\Upsilon^-$; in the case of $\Upsilon^+\Upsilon^- \geq 0$, the distribution is phase-shifted by π relative to the case of $\Upsilon^+\Upsilon^- < 0$. To incorporate this into a single consistent CP-sensitive observable, we define φ^* as

$$\varphi^* = \begin{cases} \varphi' & \text{if } \Upsilon^+\Upsilon^- \geq 0 \\ (\varphi' + \pi) \bmod 2\pi & \text{if } \Upsilon^+\Upsilon^- < 0 \end{cases}. \tag{6}$$

Before being input to the network, all feature distributions are standardised to have zero mean and unit variance. A selection of the feature distributions in the training data is shown in Fig. 1. The univariate feature distributions are severely overlapping for H and A events, indicating that the classification task is very challenging. The one feature which stands out here is φ^*, which is the basis for the single-variable mixture estimation described in Sect. 4.

For the results presented in Sect. 5 we use a network trained on all features discussed above. However, features such as φ^* and m_T are derived from the basic pion momenta that the network also has access to. These "high-level" features can in principle be inferred by the network itself from the "low-level" pion momenta. To briefly investigate this we repeat the network training with varying subsets of the input features, starting with only the pion four-momenta and sequentially adding φ^*, $\Upsilon^\pm$ and the remaining features. For all networks we obtain ROC AUC scores of ~ 0.630. While a full statistical comparison of the resulting networks is beyond the scope of our study, this indicates that the network is itself able to extract the relevant information from high-dimensional correlations between the pion momenta, making the explicit inclusion of the high-level inputs mostly redundant. We note that this observation is in agreement with the results of [1,2].

It is still interesting to investigate how much of the discriminatory power can be captured by the high-level features alone. For this we train several classifiers on high-level features only, adding a new set of features for each classifier. The first classifier is trained only on φ^* and achieves a ROC AUC score of ~ 0.605. When Υ^+ and Υ^- are included as input

features the performance improves to a score of ~ 0.618. This improvement can be understood qualitatively from the fact that the difference between the $\Upsilon^{\pm}$-conditional φ^* distributions for H and A events increases with $|\Upsilon^{\pm}|$. Adding E_T^{miss}, m_T, $\pi^{\pm}$ impact parameter vectors and $\mathscr{O}^*$ raises the ROC AUC score to ~ 0.620, and finally including m_{vis} further increases the score to ~ 0.623, which seems to be the limit for our network when trained on high-level features only. This indicates that φ^* and $\Upsilon^{\pm}$ together capture most of the sensitivity, but that the neural network is able to extract from the pion four-momenta some additional information which is not contained in the high-level quantities. Similar behaviour was seen in [19] in a study focusing on the CP-nature of the 125 GeV Higgs.

3.3 Network layout

In this study we employ a fully-connected feed-forward network. The input layer has 26 nodes, followed by 500 nodes in the first hidden layer, 1000 nodes in the second hidden layer, and 100 nodes in the final hidden layer. These have leaky ReLU [32] activation functions, and dropout [33] is applied with a dropping probability of 0.375. No further regularisation is imposed. All network weights are initialised from a normal distribution, following the He procedure [34]. The output layer has a softmax activation function, and we apply batch normalisation [35] between all layers. The weights are optimised using Adam [36] with cross-entropy loss and an initial learning rate of 0.03. 20% of the training data are set aside to validate the model performance during training. If there is no improvement of the loss on the validation data for ten consecutive epochs, the learning rate is reduced by a factor ten. The network is trained for 100 epochs or until no improvement is observed during 15 epochs, whichever occurs first. The neural network implementation is done using the Keras [37] and TensorFlow [38] frameworks.

4 The φ^* method

Traditional approaches for separating CP-even and -odd decays are based on the angle φ^* between the tau decay planes, as defined in Eq. (6). The φ^* distribution for H and A events can be seen in Fig. 2a. The CP-sensitive parameter in this distribution is the phase of the sinusoidal curve, which is shifted by π radians between the H and A hypotheses. We note that the distributions overlap across the full φ^* range, hence no absolute event separation is possible based on this variable.

Using the simplified notation $p(\varphi^*|A) \equiv p_A(\varphi^*)$ and $p(\varphi^*|H) \equiv p_H(\varphi^*)$, the φ^* distribution for H/A signal data can be expressed as a simple mixture model,

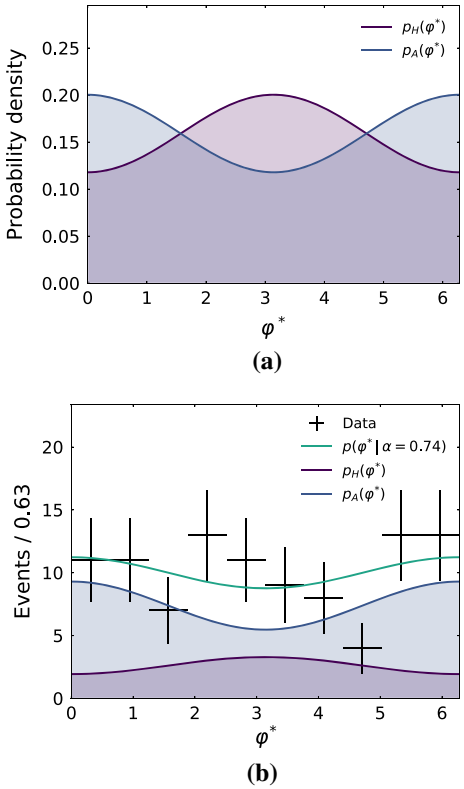


Fig. 2 **a** The probability density for φ^* in H events ($p_H(\varphi^*)$) and A events ($p_A(\varphi^*)$). **b** A fit of the mixture model $p(\varphi^*|\alpha) = \alpha p_A(\varphi^*) + (1-\alpha)p_H(\varphi^*)$ to a test dataset. Data points are shown in black, while the fitted model (normalized to 100 events) is shown in green. For this dataset the best-fit α value is $\hat{\alpha} = 0.74$

$$\begin{aligned} p(\varphi^*|\alpha) &= \alpha p_A(\varphi^*) + (1-\alpha) p_H(\varphi^*) \\ &= \alpha\big(a\cos\varphi^* + c\big) + \big(1-\alpha\big)\big(a\cos(\varphi^* + \pi) + c\big) \\ &= \alpha\, a\cos\varphi^* + (1-\alpha)\, a\cos(\varphi^* + \pi) + c, \end{aligned} \tag{7}$$

where we fix the amplitude a and offset c to $a = 0.041$ and $c = 0.159$, obtained from a separate fit to H and A training data. This leaves us with a model for the φ^* distribution where α is the only free parameter. Given a dataset $\{\varphi_i^*\}$ with N events, we can now obtain an estimate $\hat{\alpha}$ for α by maximising the likelihood function

$$\hat{\alpha} = \arg\max_{\alpha} \prod_{i=1}^{N} p(\varphi_i^*|\alpha). \tag{8}$$

We demonstrate this method in Fig. 2 for a dataset with 100 H/A Pythia events, generated using a model with a true α of 0.7. The pdfs $p_H(\varphi^*)$ and $p_A(\varphi^*)$ are shown in Fig. 2a, while the fit result is shown in Fig. 2b. For this example the best-fit α estimate comes out at $\hat{\alpha} = 0.74$.

To demonstrate the statistical performance of this estimator we repeat the fit using 10,000 independent test sets with 100 Pythia events each, generated with true α values

of 0.5, 0.7 and 0.9. The resulting distributions of α estimates are shown in Fig. 4a, where the purple (green) distributions depict results without (with) detector effects. By fitting a Gaussian to each distribution we find the spread in the estimates to be $\sigma_\alpha = 0.27$ ($\sigma_\alpha^{\text{det}} \simeq 0.45$) when detector smearing is omitted (included). Further, the estimator is mean-unbiased for all three cases. Note that to demonstrate the unbiasedness we have allowed the fit to vary α beyond the physically valid range of [0, 1].

5 The neural network method

When estimating some parameter θ using collider data we ideally want to make use of the multivariate density $p(\mathbf{x}|\theta)$ for the complete set of event features $\mathbf{x}$.[5] However, it is typically infeasible to evaluate this density directly for a given $\mathbf{x}$. A common approach is then to construct a new variable $y(\mathbf{x})$ and base the parameter estimation on the simpler, univariate distribution $p(y(\mathbf{x})|\theta)$, as exemplified by the φ^* fit in Sect. 4.

The performance of such a univariate approach depends on how well the distribution $p(y(\mathbf{x})|\theta)$ retains the sensitivity to θ found in the underlying distribution $p(\mathbf{x}|\theta)$. In the special case where the map $y(\mathbf{x})$ is the output from a trained classifier, it can be shown that using $p(y(\mathbf{x})|\theta)$ to estimate θ in the ideal limit is equivalent to using the full data distribution $p(\mathbf{x}|\theta)$. Here we briefly review this argument before applying the classifier approach to our mixture estimation problem.

After training on θ-labeled data, a classifier that minimizes a suitably chosen error function will approximate a decision function $s(\mathbf{x})$ that is a strictly monotonic function of the density ratio $p(\mathbf{x}|\theta)/p(\mathbf{x}|\theta')$ [39].[6] As shown in [40], the monotonicity of $s(\mathbf{x})$ ensures that density ratios based on the multivariate distribution $p(\mathbf{x}|\theta)$ and the univariate distribution $p(s(\mathbf{x})|\theta)$ are equivalent,

$$\frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta')} = \frac{p(s(\mathbf{x})|\theta)}{p(s(\mathbf{x})|\theta')}. \tag{9}$$

If we now take θ' to be a fixed value such that the support of $p(\mathbf{x}|\theta')$ covers the support of $p(\mathbf{x}|\theta)$,[7] the maximum likelihood estimator for θ based on $p(\mathbf{x}|\theta)$ can be rewritten as follows [40]:

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta \prod_{i=1}^{N} p(\mathbf{x}_i|\theta) \\
&= \arg\max_\theta \prod_{i=1}^{N} \frac{p(\mathbf{x}_i|\theta)}{p(\mathbf{x}_i|\theta')} \\
&= \arg\max_\theta \prod_{i=1}^{N} \frac{p(s(\mathbf{x}_i)|\theta)}{p(s(\mathbf{x}_i)|\theta')} \\
&= \arg\max_\theta \prod_{i=1}^{N} p(s(\mathbf{x}_i)|\theta).
\end{aligned} \tag{10}$$

Hence, if the classifier output $y(\mathbf{x})$ provides a reasonable approximation of $s(\mathbf{x})$ we can expect the maximum likelihood estimator based on $p(y(\mathbf{x})|\theta)$ to exhibit similar performance to an estimator based on $p(\mathbf{x}|\theta)$. The main drawbacks of this approach are the complications associated with training the classifier, and that the physics underlying the parameter sensitivity may remain hidden from view.

We now apply this classifier approach to our H/A mixture estimation problem. The maximum likelihood estimator for the mixture parameter α is then given by

$$\begin{aligned}
\hat{\alpha} &= \arg\max_\alpha \prod_{i=1}^{N} p(y(\mathbf{x}_i)|\alpha) \\
&= \arg\max_\alpha \prod_{i=1}^{N} \Big[\alpha p_A(y(\mathbf{x}_i)) + (1-\alpha) p_H(y(\mathbf{x}_i))\Big],
\end{aligned} \tag{11}$$

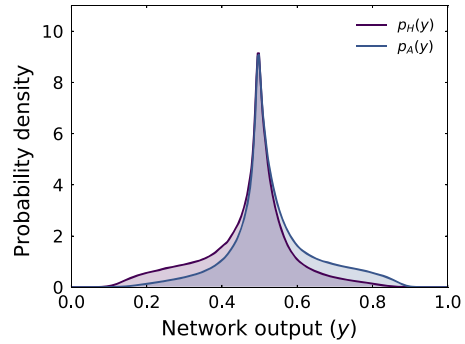
where we have expressed the overall network output distribution $p(y|\alpha)$ as a mixture of the pure-class distributions $p(y|A) \equiv p_A(y)$ and $p(y|H) \equiv p_H(y)$. We use a network trained on a balanced set of H and A events. The network is trained to associate outputs $y = 0$ and $y = 1$ with H and A events, respectively. By applying this network to another labeled dataset of equal size to the training set, we construct templates for the probability densities $p_H(y)$ and $p_A(y)$ in Eq. (11) using a nonparametric kernel density estimation method (KDE) [41]. The resulting templates are shown in Fig. 3a. We note that the pdfs do not span the entire allowed range $y \in [0, 1]$. This is expected, since the CP nature of a single event cannot be determined with complete certainty. Proper determination of the pdf shapes in the extremities – where the sensitivity is highest – requires a sufficient amount of data, which is why we devote a similarly sized data set to the template creation as to the network training.

Given a set of unlabeled data we can now estimate α by carrying out the maximization in Eq. (11) as an unbinned maximum-likelihood fit. The resulting fit to the same example dataset as used for the φ^* fit in Fig. 2b is shown in Fig. 3b. The best-fit α estimate in this case is $\hat{\alpha} = 0.67$.

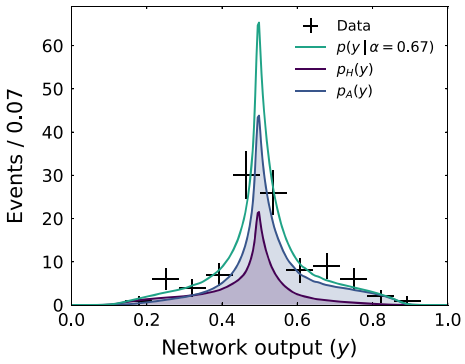
[5] Here θ represents an arbitrary model parameter, not necessarily a simple mixture parameter.

[6] In general the decision function can depend directly on the parameter values θ and θ': $s = s(\mathbf{x}; \theta, \theta')$. However, this is not the case for a mixture estimation problem like the one considered here, where $\mathbf{x}$ represents a single draw from one of the mixture model components (kinematic data from a single H or A event) and the parameter of interest is the unknown component mixture (α) of the complete dataset $\{\mathbf{x}_i\}$.

[7] This is trivially satisfied for any choice $\theta' \in (0, 1)$ when θ represents the mixture parameter of a simple two-component mixture model.



(a)



(b)

Fig. 3 **a** KDE estimate for the distribution of the network output y for H events ($p_H(y)$) and A events ($p_A(y)$), given a balanced network. **b** A fit of the mixture model $p(y|\alpha) = \alpha p_A(y) + (1-\alpha)p_H(y)$ to the same example dataset as used in Fig. 2b. Data points are shown in black and the fitted model in green. The best-fit α value is $\hat{\alpha} = 0.67$

5.1 Results

We can now compare the performance of the neural network method with that of the φ^* method of Sect. 4. To this end, we apply the network method to the same test sets as used in Fig. 4a, i.e. 10,000 datasets of 100 Pythia events each, for each of the three scenarios $\alpha = 0.5, 0.7, 0.9$. The analysis is repeated with network training and test sets with and without detector smearing. The results are given in Fig. 4b, for easy comparison with the corresponding results of the φ^* method in Fig. 4a. We fit each distribution of α estimates with a Gaussian and summarize the fit parameters in Table 1.

As for the φ^* method, we find that detector smearing significantly impacts the width of the α distribution, which increases from $\sigma_\alpha = 0.21$ to $\sigma_\alpha^{\text{det}} = 0.37$ upon inclusion of detector effects. Yet, the network approach consistently outperforms the φ^* method, as σ_α and $\sigma_\alpha^{\text{det}}$ are reduced by $\sim 22\%$ and $\sim 18\%$, respectively, compared to the φ^* results. So while the absolute widths of the α distributions in Fig. 4 illustrate that a dataset of 100 events is probably too small to obtain an accurate α estimate, the comparison with the φ^* results indicates that the relative performance gain offered by the network approach is relatively robust against detector smearing.

Similar to the φ^* method, the network method provides a mean-unbiased estimator. In order to demonstrate this we allow α to vary outside the physical range [0, 1] in our fits. However, for $\alpha > 1$, the combined mixture model $p(y|\alpha) = \alpha p_A(y) + (1-\alpha)p_H(y)$ will become negative for y values that satisfy $p_A(y)/p_H(y) < (\alpha-1)/\alpha$. This we do not allow in our fits, and in such cases we lower the α estimate until $p(y|\alpha)$ is non-negative everywhere. This choice explains the slight deviation from Gaussianity in the region around $\hat{\alpha} = 1.2$ in the bottom right plot.[8]

Figure 5 shows the distributions of α estimates for the cases of 20 events per test dataset (top row) and 500 events per test dataset (bottom row), where all sets have been generated with $\alpha = 0.7$ and no detector smearing has been included. Compared to the results with 100 events per set, σ_α for both fit methods increase (decrease) by approximately a factor $\sqrt{5}$ for the case with 20 (500) events per set, as expected from the factor 5 decrease (increase) in statistics. Thus, the relative accuracy improvement of the neural network approach over the φ^* method remains approximately the same: 30% for the 20-events case, and 25% for the 500-events case. However, the absolute spread of estimates in the 20-events case shows that this is clearly not enough statistics to obtain a useful estimate of α.

As a cross-check of the behaviour of the network fit method, we plot in Fig. 6a the distribution of the log-likelihood ratio $-2\ln(L(\alpha = 0.7)/L(\hat{\alpha}))$ for all test datasets of our benchmark point with $\alpha = 0.7$. According to Wilks' theorem [42], the distribution of this statistic should tend towards a χ^2 distribution with one degree of freedom. By overlaying a χ^2 distribution in Fig. 6a we see that this is indeed the case. Thus, confidence intervals constructed from the log-likelihood ratio for a neural network fit should have the expected coverage. In Fig. 6b we show the log-likelihood ratio curves for the example dataset used in Figs. 2b and 3b. The narrowing of the log-likelihood parabola for the network method again illustrates the increase in precision over the φ^* method.

For this study we focus only on the separation of two signal classes, not the separation of signal from background. Of course, a realistic dataset is likely to contain a significant fraction of background events. For the signal scenario studied here, the most important backgrounds are due to "fake taus" from QCD production, single Z production ($pp \to Z \to \tau\tau$), double Z and W production ($pp \to ZZ/WZ/WW \to \tau\tau + X$) and top pair production ($t\bar{t} \to WbWb \to \tau\tau + X$). While such backgrounds

[8] The same effect is not seen for the φ^* fits, as the ratio $p_A(\varphi^*)/p_H(\varphi^*) \geq 0.59$ for all φ^*, and none of the test sets prefer an α value as large as $1/(1-0.59) \approx 2.4$.

Table 1 Summary of α estimation on 10,000 independent test sets with 100 events in each set, using the φ^* fit and the neural network (NN) fit methods

True mixture parameter α	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
α Estimates (φ^* method, no detector smearing)	0.50 ± 0.27	0.71 ± 0.27	0.90 ± 0.27
α Estimates (φ^* method, with detector smearing)	0.50 ± 0.45	0.70 ± 0.46	0.90 ± 0.45
α Estimates (NN method, no detector smearing)	0.50 ± 0.21	0.70 ± 0.21	0.90 ± 0.21
α Estimates (NN method, with detector smearing)	0.48 ± 0.37	0.68 ± 0.37	0.88 ± 0.37

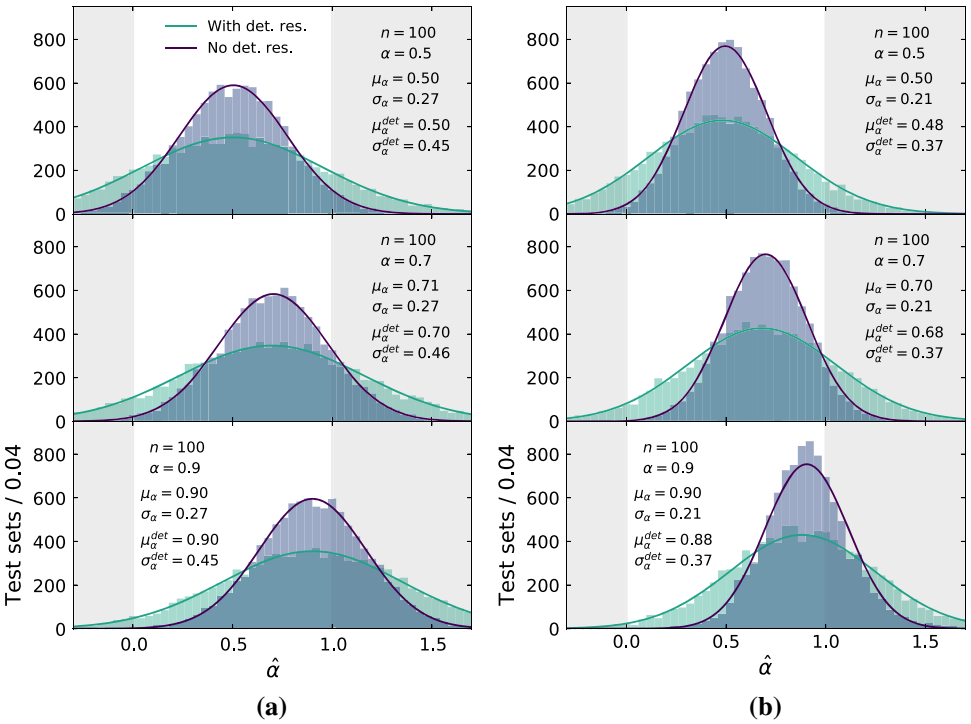


Fig. 4 Comparison of the distributions of α estimates using **a** the φ^* method and **b** the neural network method, for test sets generated with $\alpha = 0.5$ (top), $\alpha = 0.7$ (middle) and $\alpha = 0.9$ (bottom). The slight deviation from Gaussianity seen around $\hat{\alpha} = 1.2$ in the bottom right plot is due to the fact that we let α vary beyond [0, 1] in our fits, but still demand that that the mixture model $p(y|\alpha)$ is always non-negative. See the text for further details

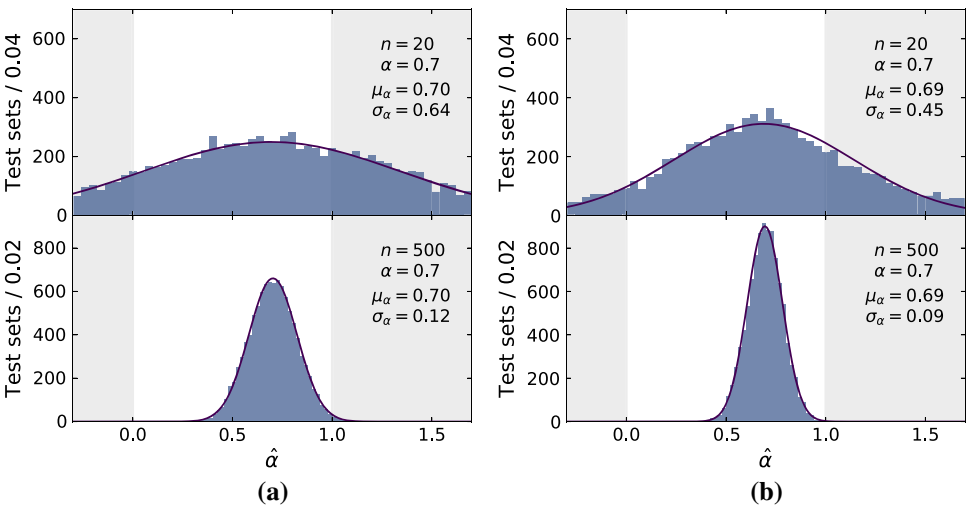
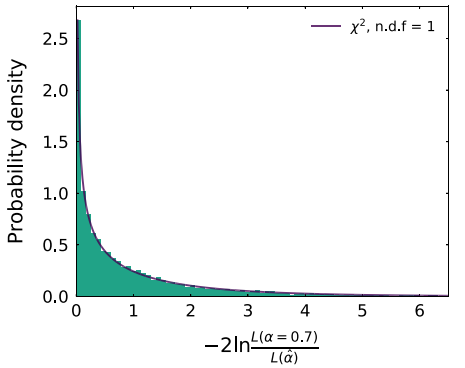
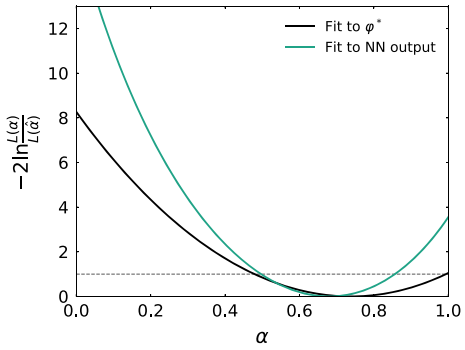


Fig. 5 Comparison of the distributions of α estimates using **a** the φ^* method and **b** the neural network method, for test sets generated with $\alpha = 0.7$. The top row shows results for test sets containing 20 events each, while the bottom row corresponds to test sets with 500 events each. The deviation from the Gaussian distribution seen at high α in the upper right plot is due to the same effect as discussed for Fig. 4

will degrade the absolute accuracy in the signal mixture estimate, it is likely to impact the φ^* method more severely than the neural network method. With one or several background components in the mixture model, the network's ability to extract information from the many-dimensional kinematic space should allow it to differentiate the background components from the signal components better than what is possible with the φ^* variable alone. We therefore expect a similar



(a)



(b)

Fig. 6 **a** Distribution of the log-likelihood ratio $-2\ln(L(\alpha = 0.7)/L(\hat{\alpha}))$ for the 10,000 test sets generated with $\alpha = 0.7$. Overlaid is a χ^2 distribution for one degree of freedom. **b** Comparison of the log-likelihood ratio curves for the test dataset from Figs. 2b and 3b, using the network method (green) and the φ^* method (black). Intersection with the horizontal dashed line at $-2\ln(L(\alpha)/L(\hat{\alpha})) = 1$ illustrates the 1σ confidence intervals, which for this example are [0.48, 1.0] for the φ^* method and [0.50, 0.86] for the neural network method

or better relative performance of the network method in the presence of background, compared to the results we have presented here. There are two ways to extend the network method to take into account additional components in the mixture model: either by implementing a multi-class classifier, or by training multiple binary classifiers on pairwise combinations of the model components. Based on [43] we expect the latter approach would give the best performance.

6 Conclusions

Estimating the component weights in mixture models with largely overlapping kinematics is a generic problem in high-energy physics. In this paper we have investigated how a deep neural network approach can improve signal mixture estimates in the challenging scenario of a ditau LHC signal coming from a pair of heavy, degenerate Higgs bosons of opposite CP charge. This is a theoretically well-motivated scenario within both general and more constrained Two-Higgs-Doublet Models.

We have studied a benchmark scenario with degenerate H and A states at $m_H = m_A = 450\,\text{GeV}$. For this case we find that the neural network approach provides a $\sim 20\%$ reduction in the uncertainty of signal mixture estimates, compared to estimates based on fitting the single most discriminating kinematic variable (φ^*). However, the improved accuracy of the neural network approach comes with a greater computational complexity.

The network method we have studied here can be extended to include additional mixture components, such as one or several background processes, either by training a multi-class classifier or by training multiple binary classifiers. To increase the available statistics, the method can also be extended to work with a wider range of tau decay modes, for instance by using the "impact parameter method" described in [18].

The code used to generate events, train the network and run the maximum likelihood estimates will be made available on gitlab.com/BSML after publication.

Acknowledgements We would like to thank Andrey Ustyuzhanin and Maxim Borisyak for helpful discussions during MLHEP 2017, and Kyle Cranmer for comments and discussions during Spåtind 2018. We also thank Ørjan Dale for helpful comments. S.M. and I.S. thank the Theory Section at the Department of Physics, University of Oslo, for the kind hospitality during the completion of this work. This work was supported by the Research Council of Norway through the FRIPRO grant 230546/F20.

Appendix A Supplementary figures

A simple scan of the high-mass parameter regions of the (SM-aligned) lepton-specific and type-I THDMs is performed to illustrate the parameter dependence of the ditau signal strengths $\sigma(pp \to H) \times \mathcal{B}(H \to \tau\tau)$ and $\sigma(pp \to A) \times \mathcal{B}(A \to \tau\tau)$, as well as the mixture parameter α. The results are shown in Fig. 7. The parameters $m_H = m_A = m_{H^\pm}$, $\tan\beta$ and m_{12}^2 are varied in the scan, while we fix the light Higgs mass $m_h = 125\,\text{GeV}$ and the neutral scalar mixing parameter $\sin(\beta - \alpha') = 1$ to ensure perfect SM alignment for the light state h. The NLO cross sections are calculated with SusHi 1.6.1, while branching ratios are calculated using 2HDMC 1.7.0. We test the parameter points against constraints from the various collider searches for Higgs bosons using HiggsBounds 4.3.1 [44–48], while the-

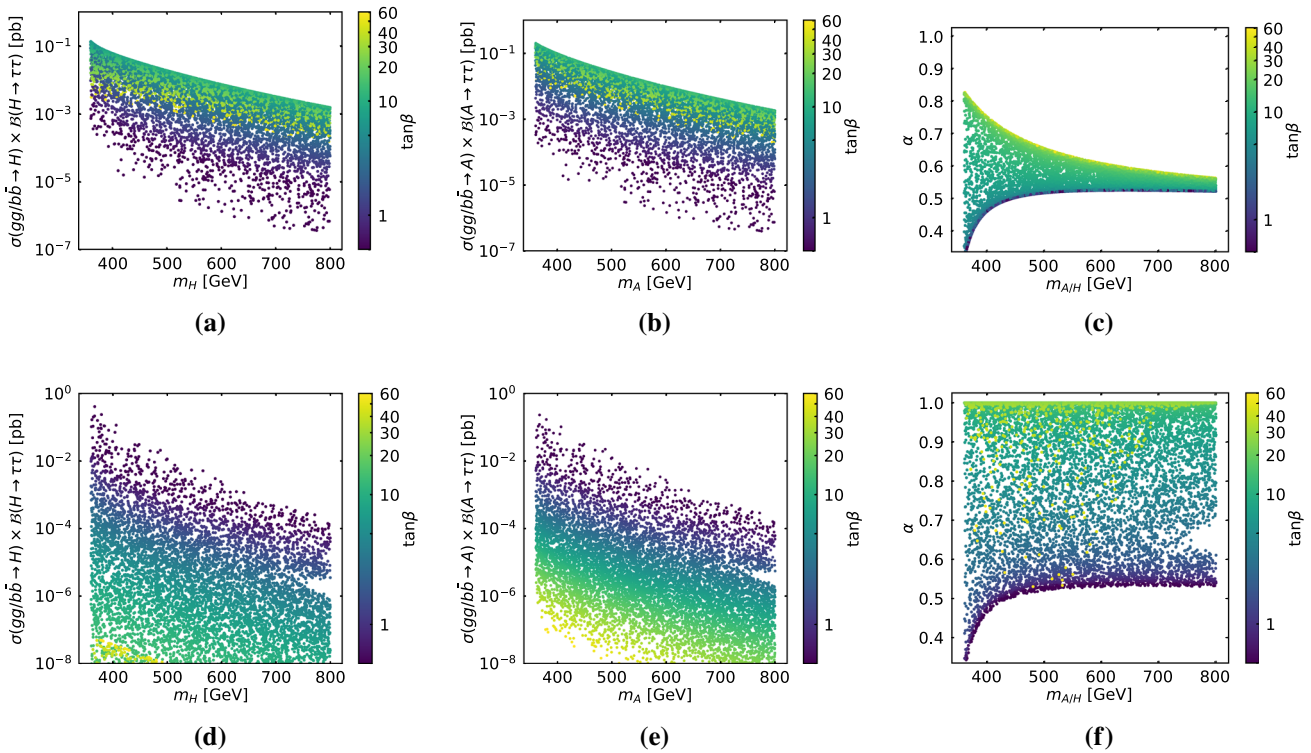


Fig. 7 *Top row, lepton-specific THDM:* **a** Signal strength for $pp \to H \to \tau\tau$, as a function of m_H and $\tan\beta$. **b** Similar result for $pp \to A \to \tau\tau$. **c** The ratio of the signal strength of $pp \to A \to \tau\tau$ to the total ditau signal strength, as defined in Eq. (1). *Bottom row:* Corresponding results within the type-I THDM

oretical constraints are checked with 2HDMC. Constraints from flavour physics, in particular $\mathcal{B}(b \to s\gamma)$, disfavour parameter regions at very low $\tan\beta$ in the type-I and lepton-specific THDMs. These constraints were not included in the simple scan.

References

1. P. Baldi, P. Sadowski, D. Whiteson, Nat. Commun. **5**, 4308 (2014). https://doi.org/10.1038/ncomms5308
2. P. Baldi, P. Sadowski, D. Whiteson, Phys. Rev. Lett. **114**(11), 111801 (2015). https://doi.org/10.1103/PhysRevLett.114.111801
3. A. Farbin, PoS **ICHEP2016**, 180 (2016)
4. V.A. Kuzmin, V.A. Rubakov, M.E. Shaposhnikov, Phys. Lett. B **155**, 36 (1985). https://doi.org/10.1016/0370-2693(85)91028-7
5. G.W. Anderson, L.J. Hall, Phys. Rev. D **45**, 2685 (1992). https://doi.org/10.1103/PhysRevD.45.2685
6. N. Turok, J. Zadrozny, Nucl. Phys. B **358**, 471 (1991). https://doi.org/10.1016/0550-3213(91)90356-3
7. N. Turok, J. Zadrozny, Nucl. Phys. B **369**, 729 (1992). https://doi.org/10.1016/0550-3213(92)90284-I
8. J.O. Andersen, T. Gorda, A. Helset, L. Niemi, T.V.I. Tenkanen, A. Tranberg, A. Vuorinen, D.J. Weir, Phys. Rev. Lett. **121**(19), 191802 (2018). https://doi.org/10.1103/PhysRevLett.121.191802
9. M. Aaboud, JHEP **01**, 055 (2018). https://doi.org/10.1007/JHEP01(2018)055
10. A.M. Sirunyan, JHEP **09**, 007 (2018). https://doi.org/10.1007/JHEP09(2018)007
11. J. Bernon, J.F. Gunion, H.E. Haber, Y. Jiang, S. Kraml, Phys. Rev. D **92**(7), 075004 (2015). https://doi.org/10.1103/PhysRevD.92.075004
12. A. Haarr, A. Kvellestad, T.C. Petersen (2016) preprint. arXiv:1611.05757
13. S.P. Martin (1997). https://doi.org/10.1142/9789812839657_0001. https://doi.org/10.1142/9789814307505_0001. [Adv. Ser. Direct. High Energy Phys. **18**, 1 (1998)]
14. P.S. Bhupal Dev, A. Pilaftsis, JHEP **12**, 024 (2014). https://doi.org/10.1007/JHEP11(2015)147. https://doi.org/10.1007/JHEP12(2014)024. [Erratum: JHEP **11**, 147 (2015)]
15. J.R. Dell'Aquila, C.A. Nelson, Nucl. Phys. B **320**(1), 61 (1989). https://doi.org/10.1016/0550-3213(89)90211-3
16. M. Kramer, J.H. Kuhn, M.L. Stong, P.M. Zerwas, Z. Phys. C **64**, 21 (1994). https://doi.org/10.1007/BF01557231
17. G.R. Bower, T. Pierzchala, Z. Was, M. Worek, Phys. Lett. B **543**, 227 (2002). https://doi.org/10.1016/S0370-2693(02)02445-0
18. S. Berge, W. Bernreuther, S. Kirchner, Phys. Rev. D **92**, 096012 (2015). https://doi.org/10.1103/PhysRevD.92.096012
19. R. Józefowicz, E. Richter-Was, Z. Was, Phys. Rev. D **94**(9), 093001 (2016). https://doi.org/10.1103/PhysRevD.94.093001
20. E. Barberio, B. Le, E. Richter-Was, Z. Was, D. Zanzi, J. Zaremba, Phys. Rev. D **96**(7), 073002 (2017). https://doi.org/10.1103/PhysRevD.96.073002
21. R.V. Harlander, S. Liebler, H. Mantler, Comput. Phys. Commun. **184**, 1605 (2013). https://doi.org/10.1016/j.cpc.2013.02.006

22. R.V. Harlander, S. Liebler, H. Mantler, Comput. Phys. Commun. **212**, 239 (2017). https://doi.org/10.1016/j.cpc.2016.10.015
23. R.V. Harlander, W.B. Kilgore, Phys. Rev. Lett. **88**, 201801 (2002). https://doi.org/10.1103/PhysRevLett.88.201801
24. R.V. Harlander, W.B. Kilgore, Phys. Rev. D **68**, 013001 (2003). https://doi.org/10.1103/PhysRevD.68.013001
25. S. Actis, G. Passarino, C. Sturm, S. Uccirati, Phys. Lett. B **670**, 12 (2008). https://doi.org/10.1016/j.physletb.2008.10.018
26. R. Harlander, P. Kant, JHEP **12**, 015 (2005). https://doi.org/10.1088/1126-6708/2005/12/015
27. K.G. Chetyrkin, J.H. Kuhn, M. Steinhauser, Comput. Phys. Commun. **133**, 43 (2000). https://doi.org/10.1016/S0010-4655(00)00155-7
28. D. Eriksson, J. Rathsman, O. Stal, Comput. Phys. Commun. **181**, 189 (2010). https://doi.org/10.1016/j.cpc.2009.09.011
29. T. Sjostrand, S. Mrenna, P.Z. Skands, JHEP **05**, 026 (2006). https://doi.org/10.1088/1126-6708/2006/05/026
30. T. Sjostrand, S. Mrenna, P.Z. Skands, Comput. Phys. Commun. **178**, 852 (2008). https://doi.org/10.1016/j.cpc.2008.01.036
31. C. Patrignani, Chin. Phys. C **40**(10), 100001 (2016). https://doi.org/10.1088/1674-1137/40/10/100001
32. A.L. Maas, A.Y. Hannun, A.Y. Ng, in *International conference on machine learning*, vol. 30, p. 3 (2013)
33. N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, J. Mach. Learn. Res. **15**(1), 1929 (2014)
34. K. He, X. Zhang, S. Ren, J. Sun, CoRR arXiv:1502.01852 (2015)
35. S. Ioffe, C. Szegedy, CoRR arXiv:1502.03167 (2015)
36. D.P. Kingma, J. Ba, CoRR arXiv:1412.6980 (2014)
37. F. Chollet, et al. Keras. https://github.com/fchollet/keras (2015)
38. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from http://tensorflow.org/
39. J. Friedman, T. Hastie, R. Tibshirani, Ann. Stat. **28**(2), 337 (2000). https://doi.org/10.1214/aos/1016218223
40. K. Cranmer, J. Pavez, G. Louppe (2015) preprint. arXiv:1506.02169
41. K.S. Cranmer, Comput. Phys. Commun. **136**, 198 (2001). https://doi.org/10.1016/S0010-4655(00)00243-5
42. S.S. Wilks, Ann. Math. Stat. **9**(1), 60 (1938). https://doi.org/10.1214/aoms/1177732360
43. K. Cranmer, J. Pavez, G. Louppe, W.K. Brooks, J. Phys. Conf. Ser. **762**(1), 012034 (2016). https://doi.org/10.1088/1742-6596/762/1/012034
44. P. Bechtle, O. Brein, S. Heinemeyer, G. Weiglein, K.E. Williams, Comput. Phys. Commun. **181**, 138 (2010). https://doi.org/10.1016/j.cpc.2009.09.003
45. P. Bechtle, O. Brein, S. Heinemeyer, G. Weiglein, K.E. Williams, Comput. Phys. Commun. **182**, 2605 (2011). https://doi.org/10.1016/j.cpc.2011.07.015
46. P. Bechtle, O. Brein, S. Heinemeyer, O. Stal, T. Stefaniak, G. Weiglein, K. Williams, PoS **CHARGED2012**, 024 (2012)
47. P. Bechtle, O. Brein, S. Heinemeyer, O. Stal, T. Stefaniak, G. Weiglein, K.E. Williams, Eur. Phys. J. C **74**(3), 2693 (2014). https://doi.org/10.1140/epjc/s10052-013-2693-2
48. P. Bechtle, S. Heinemeyer, O. Stal, T. Stefaniak, G. Weiglein, Eur. Phys. J. C **75**(9), 421 (2015). https://doi.org/10.1140/epjc/s10052-015-3650-z