

DNN approach: Higgs CP mixing angle in $H \rightarrow \tau\tau$ decay

E. Richter-Was (IF UJ, Krakow)

- DNN tasks
- Few formulas from theory
- Feature lists
- Results for classification and regression
- Outlook
- Additional slides: DNN details

MC samples:

Events of ggH , generated with Pythia8, decayed with Tauola with polarisation OFF. Spin weight calculated with TauSpinner. About 10M events for each decay channels.

Presented results from:

R. Jozefowicz et al. (ERW), Phys. Rev. D94 (2016) 093001, arXiv: 1608.02609

E. Barberio et al. (B. Le, ERW, D. Zanzi), Phys. Rev. D96 (2017) 073002, arXiv: 1706.07983

K. Lasocha et al. (ERW), Phys. Rev. D100 (2019) 113001, arXiv: 1812.08140

K. Lasocha et al. (ERW), Phys. Rev. D103 (2021) 036003, arXiv: 200100455

DNN tasks: determine per event parameters

Goals are complementary and to large extend also redundant.

DNN is trained to predict **per event**:

- Probability that event is of class A given alternative hypotheses
- Spin weight as a function of the CP mixing angle
- Decay configuration dependent coefficients of spin weight formula sensitive to CP mixing angle
- The most preferred mixing angle

Methods:

- **Binary classification:** discriminating between two hypotheses
- **Multiclass classification:** simultaneously learning probabilities for several hypotheses
- **Regression:** learning values associated with the event


CP sensitive term in lagrangian

The most general Higgs boson Yukawa coupling, expressed with help of scalar-pseudoscalar parity mixing angle ϕ^{CP} reads

$$\mathcal{L}_Y = N\bar{\tau}h(\cos\phi^{CP} + i\sin\phi^{CP}\gamma_5)\tau, \quad \phi^{CP} \text{ spans over } (0,\pi) \text{ range}$$

Starting from this lagrangian, the squared matrix element term

polarimetric vector **correlation matrix**

$$|M|^2 \sim 1 + h_+^i h_-^j R_{i,j}; \quad i, j = \{x, y, z\}$$


The corresponding CP sensitive weight

$$wt = 1 - h_+^z h_-^z + h_+^\perp R(2\phi^{CP}) h_-^\perp.$$

$$R_{xx} = R_{yy} = \cos 2\phi^{CP}, \quad R_{xy} = -R_{yx} = \sin 2\phi^{CP}$$

CP sensitive spin weight

For clarity of results we introduced

$$\alpha^{CP} = 2\phi^{CP}, \text{ which spans over } (0, 2\pi) \text{ range.}$$

$\alpha^{CP} = 0, 2\pi$	- scalar
$\alpha^{CP} = \pi$	- pseudoscalar

The CP sensitive weight

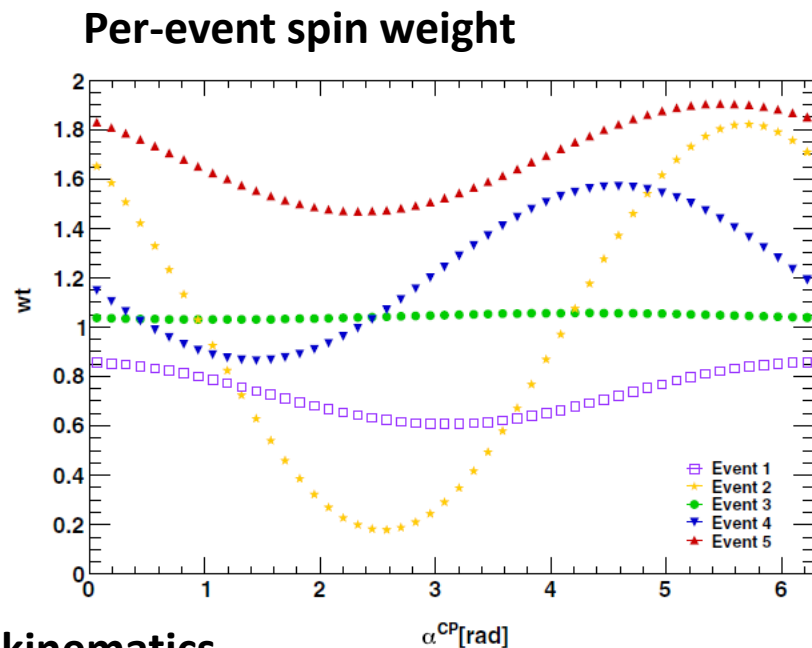
$$wt = C_0 + C_1 \cdot \cos(\alpha^{CP}) + C_2 \cdot \sin(\alpha^{CP}),$$

$$C_0 = 1 - h_+^z h_-^z,$$

$$C_1 = -h_+^x h_-^x + h_+^y h_-^y,$$

$$C_2 = -h_+^x h_-^y - h_+^y h_-^x,$$

C_i coefficients depend only on the tau decay kinematics,
they are products of polarimetric vectors.



CP sensitive spin weight

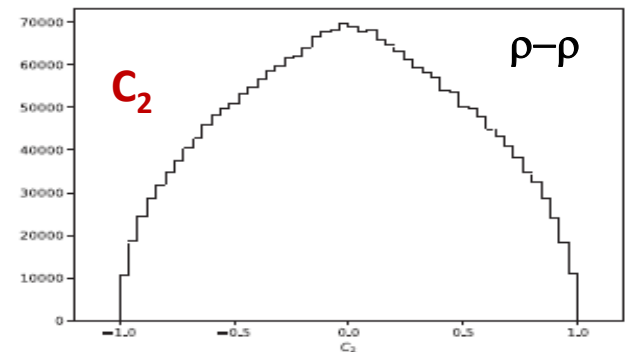
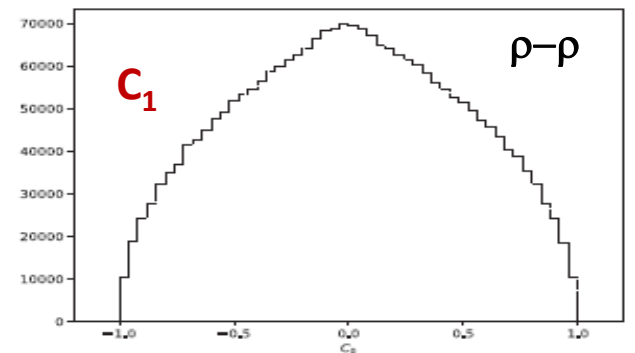
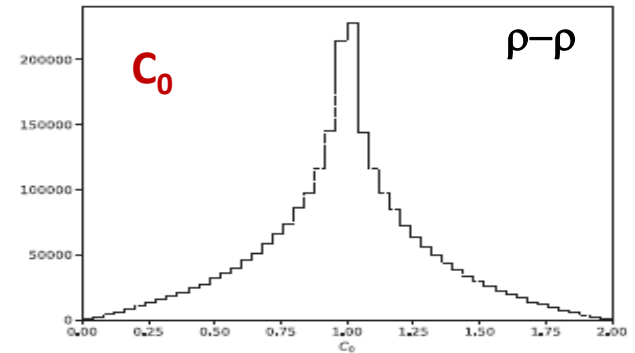
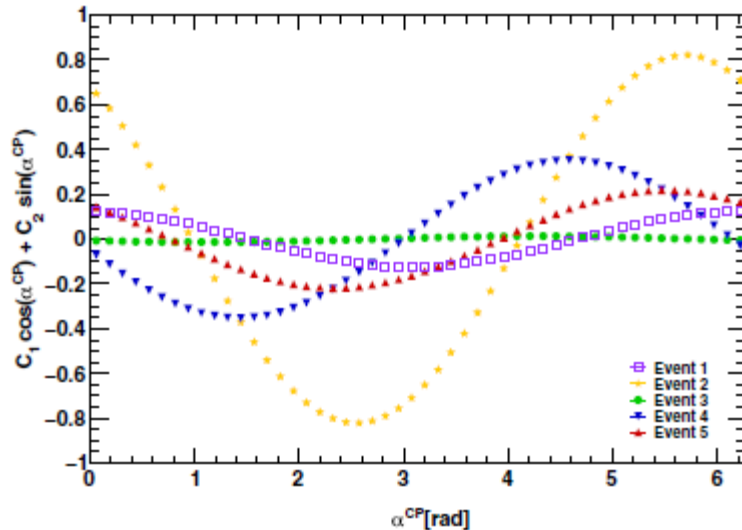
$$wt = C_0 + \underline{C_1 \cdot \cos(\alpha^{CP}) + C_2 \cdot \sin(\alpha^{CP})},$$

$$C_0 = 1 - h_+^z h_-^z,$$

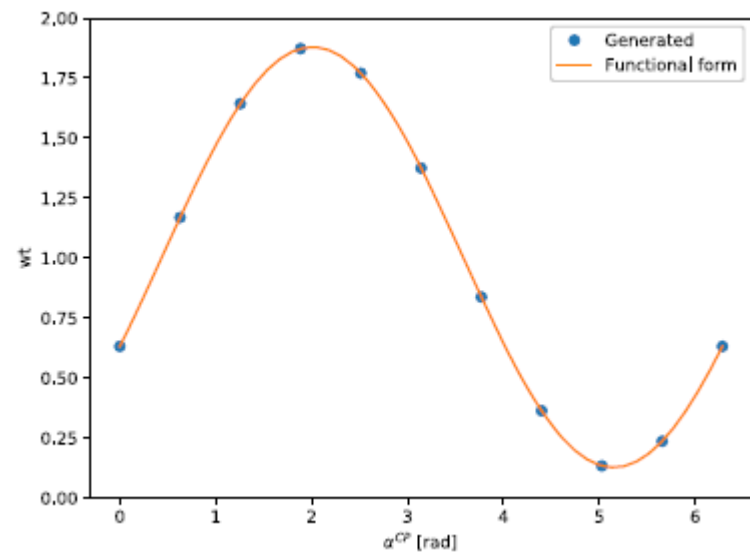
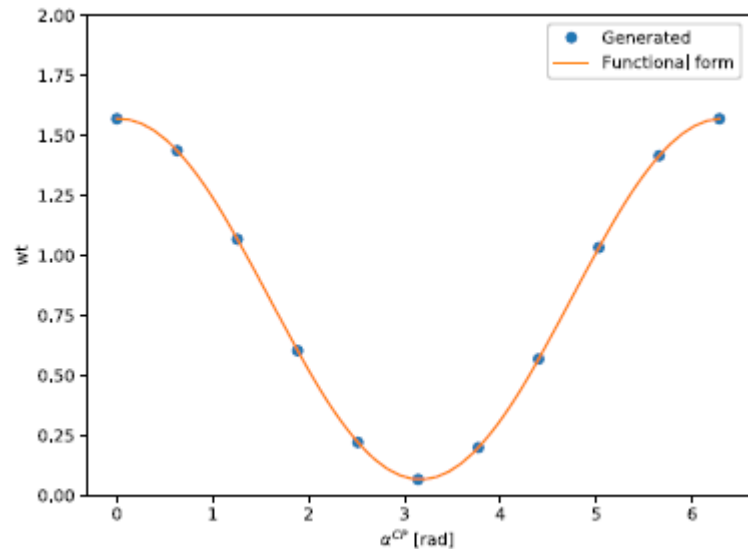
$$C_1 = -h_+^x h_-^x + h_+^y h_-^y,$$

$$C_2 = -h_+^x h_-^y - h_+^y h_-^x,$$

Only CP dependent part of the spin weight



Cross-check: MC generated vs functional form



Closure test for implementation in the MC:
calculated per-event C_i , compared with functional form and spin weight of TauSpinner

Feature lists and binary classification

Phys. Rev. D94 (2016) 093001

Features/variables	Decay mode: $\rho^\pm - \rho^\mp$ $\rho^\pm \rightarrow \pi^0 \pi^\pm$	Decay mode: $a_1^\pm - \rho^\mp$ $a_1^\pm \rightarrow \rho^0 \pi^\pm, \rho^0 \rightarrow \pi^+ \pi^-$ $\rho^\mp \rightarrow \pi^0 \pi^\mp$	Decay mode: $a_1^\pm - a_1^\mp$ $a_1^\pm \rightarrow \rho^0 \pi^\pm, \rho^0 \rightarrow \pi^+ \pi^-$
$\varphi_{i,k}^*$	1	4	16
$\varphi_{i,k}^*$ and y_i, y_k	3	9	24
$\varphi_{i,k}^*$, 4-vectors	25	36	64
$\varphi_{i,k}^*, y_i, y_k$ and m_i, m_k	5	13	30
$\varphi_{i,k}^*, y_i, y_k, m_i, m_k$ and 4-vectors	29	45	78

high level

high and low level

Table 2: Dimensionality of the features which may be used in each discussed configuration of the decay modes. Note that in principle y_i^\pm, y_k^\mp may be calculated in the rest frame of the resonance pair used to define $\varphi_{i,k}^*$ planes, but in practice, choice of the frames is of no numerically significant effect. We do not distinguish such variants.

Features/variables	Decay mode: $\rho^\pm - \rho^\mp$ $\rho^\pm \rightarrow \pi^0 \pi^\pm$	Decay mode: $a_1^\pm - \rho^\mp$ $a_1^\pm \rightarrow \rho^0 \pi^\mp, \rho^0 \rightarrow \pi^+ \pi^-$ $\rho^\mp \rightarrow \pi^0 \pi^\mp$	Decay mode: $a_1^\pm - a_1^\mp$ $a_1^\pm \rightarrow \rho^0 \pi^\pm, \rho^0 \rightarrow \pi^+ \pi^-$
True classification	0.782	0.782	0.782
$\varphi_{i,k}^*$	0.500	0.500	0.500
$\varphi_{i,k}^*$ and y_i, y_k	0.624	0.569	0.536
4-vectors	0.638	0.590	0.557
$\varphi_{i,k}^*$, 4-vectors	0.638	0.594	0.573
$\varphi_{i,k}^*, y_i, y_k$ and m_i^2, m_k^2	0.626	0.578	0.548
$\varphi_{i,k}^*, y_i, y_k, m_i^2, m_k^2$ and 4-vectors	0.639	0.596	0.573

Oracle prediction

Table 3: Average probability p_i that a model predicts correctly event x_i to be of a type A (scalar), with training being performed for separation between type A and B (pseudo-scalar).

Binary classification and syst. from detector resolution

Smearing: gaussian smearing of the 4-vectors

Phys. Rev. D96 (2017) 073002

Charged [16]: $\sigma(\theta) = 0.88$ mrad, $\sigma(\phi) = 0.147$ mrad and $\sigma(1/p) = 4.83 \times 10^{-4} \text{ GeV}^{-1}$

Neutral [17]: $\sigma(\eta) = 0.0056$ rad, $\sigma(\phi) = 0.012$ rad and $\sigma(E_T) = 0.16 \cdot E_T$

Features				Exact \pm (stat)	Smeared \pm (stat) \pm (syst)	From [11]
ϕ^*	4-vec	y_i	m_i			
$a_1 - \rho$ Decays						
✓	✓	✓	✓	0.6035 ± 0.0005	$0.5923 \pm 0.0005 \pm 0.0002$	0.596
✓	✓	✓	-	0.5965 ± 0.0005	$0.5889 \pm 0.0005 \pm 0.0002$	-
✓	✓	-	✓	0.6037 ± 0.0005	$0.5933 \pm 0.0005 \pm 0.0003$	-
-	✓	-	-	0.5971 ± 0.0005	$0.5892 \pm 0.0005 \pm 0.0002$	0.590
✓	✓	-	-	0.5971 ± 0.0005	$0.5893 \pm 0.0005 \pm 0.0002$	0.594
✓	-	✓	✓	0.5927 ± 0.0005	$0.5847 \pm 0.0005 \pm 0.0002$	0.578
✓	-	✓	-	0.5819 ± 0.0005	$0.5746 \pm 0.0005 \pm 0.0002$	0.569
$a_1 - a_1$ Decays						
✓	✓	✓	✓	0.5669 ± 0.0004	$0.5657 \pm 0.0004 \pm 0.0001$	0.573
✓	✓	✓	-	0.5596 ± 0.0004	$0.5599 \pm 0.0004 \pm 0.0001$	-
✓	✓	-	✓	0.5677 ± 0.0004	$0.5661 \pm 0.0004 \pm 0.0001$	-
-	✓	-	-	0.5654 ± 0.0004	$0.5641 \pm 0.0004 \pm 0.0001$	0.553
✓	✓	-	-	0.5623 ± 0.0004	$0.5615 \pm 0.0004 \pm 0.0001$	0.573
✓	-	✓	✓	0.5469 ± 0.0004	$0.5466 \pm 0.0004 \pm 0.0001$	0.548
✓	-	✓	-	0.5369 ± 0.0004	$0.5374 \pm 0.0004 \pm 0.0001$	0.536

Table 1. AUC for NN trained to separate scalar and pseudoscalar hypotheses with combinations of input features marked with a ✓. Results in the column labelled “Exact” are from NNs trained with exact sample. The results in column labelled “Smeared” are from NNs trained with smeared sample. Statistical uncertainties are derived from a bootstrap method described in the text. Systematic uncertainty is calculated with the method described in the text.

[11] R. Józefowicz, E. Richter-Was, and Z. Was, Phys. Rev. **D94**, 093001 (2016), 1608.02609.

[16] ATLAS, G. Aad *et al.*, Eur. Phys. J. **C70**, 787 (2010), 1004.5293.

[17] ATLAS, G. Aad *et al.*, Eur. Phys. J. **C76**, 295 (2016), 1512.05955.

Binary classification and syst. from τ decay model

Phys. Rev. D96 (2017) 073002

Features				STD	$R_{\chi L}$	ALT	BBR
ϕ^*	4-vec	y_i	m_i				
$a_1 - \rho$ Decays							
✓	✓	✓	✓	0.604	0.604	0.603	0.603
✓	✓	✓	-	0.597	0.596	0.596	0.597
✓	✓	-	✓	0.604	0.604	0.604	0.604
-	✓	-	-	0.597	0.596	0.596	0.595
✓	✓	-	-	0.597	0.596	0.596	0.595
✓	-	✓	✓	0.593	0.593	0.593	0.593
✓	-	✓	-	0.582	0.579	0.580	0.578
$a_1 - a_1$ Decays							
✓	✓	✓	✓	0.567	0.563	0.564	0.564
✓	✓	✓	-	0.560	0.555	0.557	0.556
✓	✓	-	✓	0.568	0.564	0.566	0.566
-	✓	-	-	0.562	0.557	0.559	0.559
✓	✓	-	-	0.562	0.557	0.559	0.559
✓	-	✓	✓	0.547	0.546	0.547	0.545
✓	-	✓	-	0.537	0.534	0.535	0.533

Table 2. Area under ROC curve. NN trained with $a_1 - a_1$ decays of $\tau\tau$ system with standard CLEO current on exact MC sample. These NN are then tested on MC generated sample with alternative parameterisations of the hadronic currents.

STD, $R_{\chi L}$, ALT, BBR – models for a_1 decay

Binary classification and feature lists

Phys. Rev. D100 (2019) 113001

TABLE II. Lists of features for ML classification, marked as Variant-X.Y. In the third column, the number of features for the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels are given. All components of the 4-momenta are taken in the hadronic decay product rest frame. The primary resonances (ρ^\pm , a_1^\pm) are aligned with the z axis. E_{miss}^x and E_{miss}^y are in the laboratory frame. In practice, instead of p_ν^T and ϕ_ν , the pair of variables $p_\nu^T \cos \phi_\nu$ and $p_\nu^T \sin \phi_\nu$ is used.

Notation	Features	Counts	Comments
Variant-All	4-momenta (π^\pm , π^0 , ν)	24/28/32	
Variant-1.0	4-momenta (π^\pm , π^0)	16/20/24	as in Table 3 of Ref. [20]
Variant-1.1	4-momenta (π^\pm , π^0 , ρ^\pm , a_1^\pm), m_i^2 , m_k^2 , y_i , y_k , $\phi_{i,k}^*$	29/46/94	
Variant-2.0	4-momenta (π^\pm , π^0), E_ν , p_ν^z , p_ν^T	22/26/30	
Variant-2.1	4-momenta (π^\pm , π^0), E_ν , p_ν^z , p_ν^T	22/26/30	Approx. E_ν , p_ν^z , p_ν^T
Variant-2.2	4-momenta (π^\pm , π^0), E_ν , p_ν^z , p_ν^T , E_{miss}^x , E_{miss}^y	24/28/32	Approx. E_ν^z , p_ν , p_ν^T
Variant-3.0.0	4-momenta (π^\pm , π^0), E_ν , p_ν^z , p_ν^T , ϕ_ν	24/28/32	Approx. E_ν , \vec{p}_ν
Variant-3.1. β	4-momenta (π^\pm , π^0), E_ν , p_ν^z , p_ν^T , ϕ_ν	24/28/32	Approx. E_ν , \vec{p}_ν ; ϕ_ν smeared with β
Variant-4.0	4-momenta (π^\pm , π^0 , τ^\pm)	24/28/32	
Variant-4.1	4-momenta (π^\pm , π^0 , τ^\pm)	24/28/32	Approx. p_τ

Binary classification and feature lists

AUC – Area Under ROC Curve

APS - average precision

Phys. Rev. D100 (2019) 113001

TABLE III. The AUC and APS scores to discriminate scalar and pseudoscalar CP states of the Higgs boson, obtained on the test sample. The DNN was trained on 50 epochs with a dropout of 0.2 (except for the explicitly marked case of Variant-All). Results for the $\rho^\pm - \rho^\mp$, $a_1^\pm - \rho^\mp$, and $a_1^\pm - a_1^\mp$ channels are given. The first column labels the choice of features. For details, see Table II.

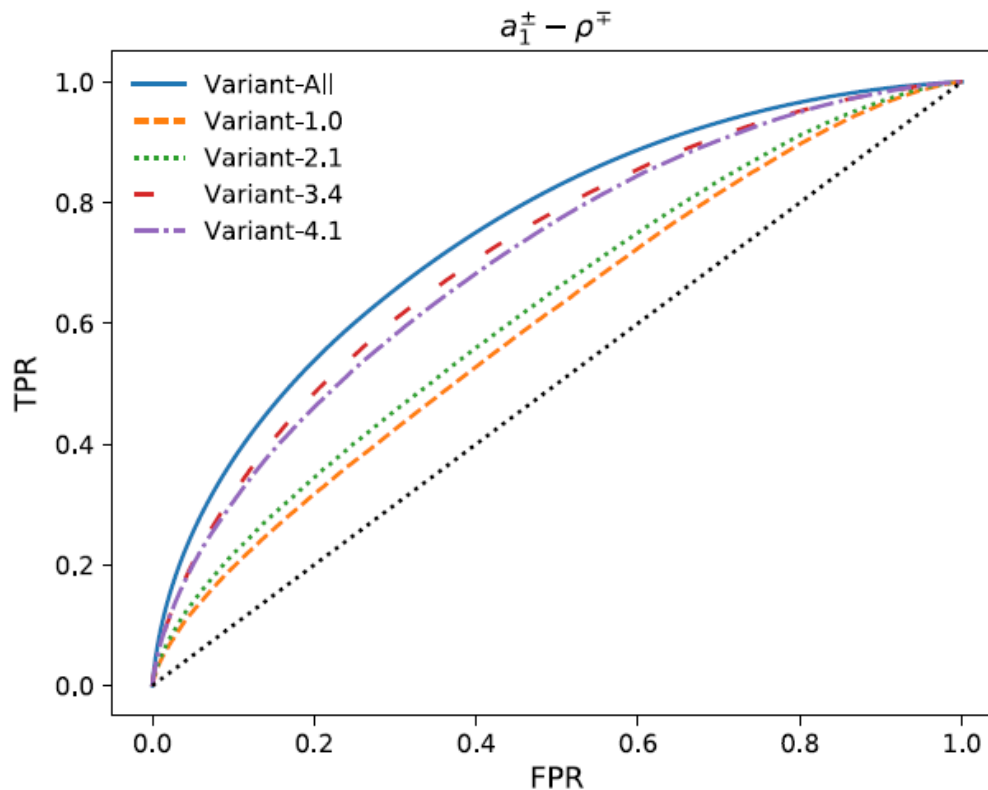
Features list	AUC/APS ($\rho^\pm - \rho^\mp$)	AUC/APS ($a_1^\pm - \rho^\mp$)	AUC/APS ($a_1^\pm - a_1^\mp$)
Oracle predictions	0.784/0.785	0.781/0.783	0.780/0.782
Variant-All (drop = 0.0)	0.784/0.786	0.778/0.778	0.773/0.774
Variant-All	0.769/0.764	0.748/0.742	0.728/0.720
Variant-1.0	0.655/0.654	0.603/0.602	0.573/0.578
Variant-1.1	0.656/0.655	0.609/0.607	0.580/0.585
Variant-2.0	0.663/0.663	0.626/0.625	0.594/0.595
Variant-2.1	0.664/0.666	0.622/0.622	0.591/0.593
Variant-2.2	0.664/0.666	0.622/0.622	0.591/0.593
Variant-3.0.0	0.771/0.771	0.749/0.743	0.728/0.721
Variant-3.1.2	0.760/0.759	0.738/0.730	0.718/0.710
Variant-3.1.4	0.738/0.735	0.714/0.705	0.687/0.677
Variant-3.1.6	0.715/0.713	0.689/0.680	0.660/0.652
Variant-4.0	0.769/0.766	0.748/0.742	0.728/0.720
Variant-4.1	0.738/0.733	0.704/0.696	0.683/0.676

Binary classification and feature lists

Phys. Rev. D100 (2019) 113001

TPR - true positive rates

FPR - false positive rates



The rest of presentation is based on **Phys. Rev. D103 (2021) 036003**

- ρ – ρ channel
- **Variant-All** for feature list, i.e. 4-momenta of π^\pm , π^0 , ν

Notation	Features	Counts
Variant-All	4-momenta (π^\pm , π^0 , ν)	24

Binary classification and CP mixing

Oracle prediction= Bayes optimal probability
sampled from hypothesis \mathcal{H}_0 vs $\mathcal{H}_{\alpha^{CP}}$
(true probability)

\mathcal{H}_0 - Higgs coupling is as for a scalar

$\mathcal{H}_{\alpha^{CP}}$ - Higgs coupling is a mixture of scalar
and pseudoscalar with mixing angle α^{CP}

CP-mixing angle α^{CP} (units of π)	Oracle predictions	Binary classification
0.2	0.528	0.525
0.4	0.605	0.595
0.6	0.699	0.684
0.8	0.775	0.756
1.0	0.804	0.784

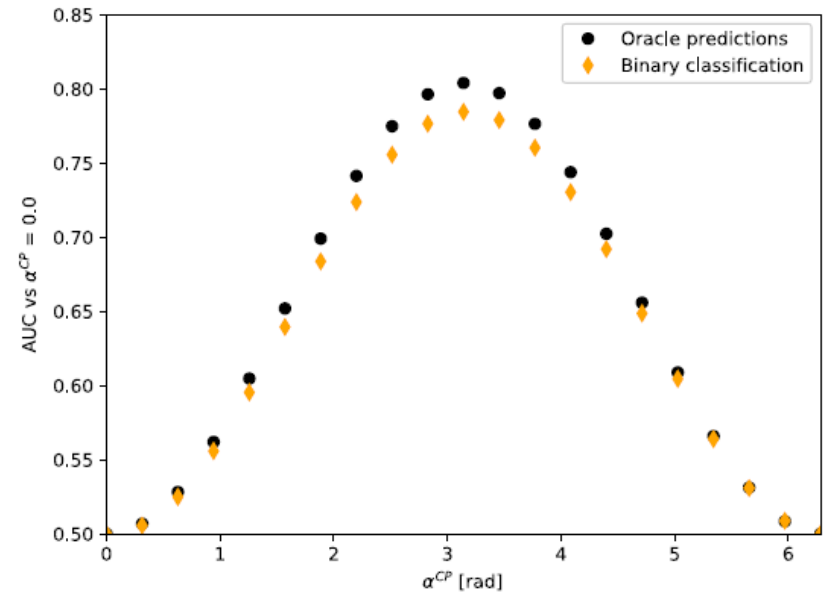


FIG. 4. The AUC score for binary classification between \mathcal{H}_0 and $\mathcal{H}_{\alpha^{CP}}$ hypotheses and corresponding oracle predictions.

Multiclass classification: learning spin weight

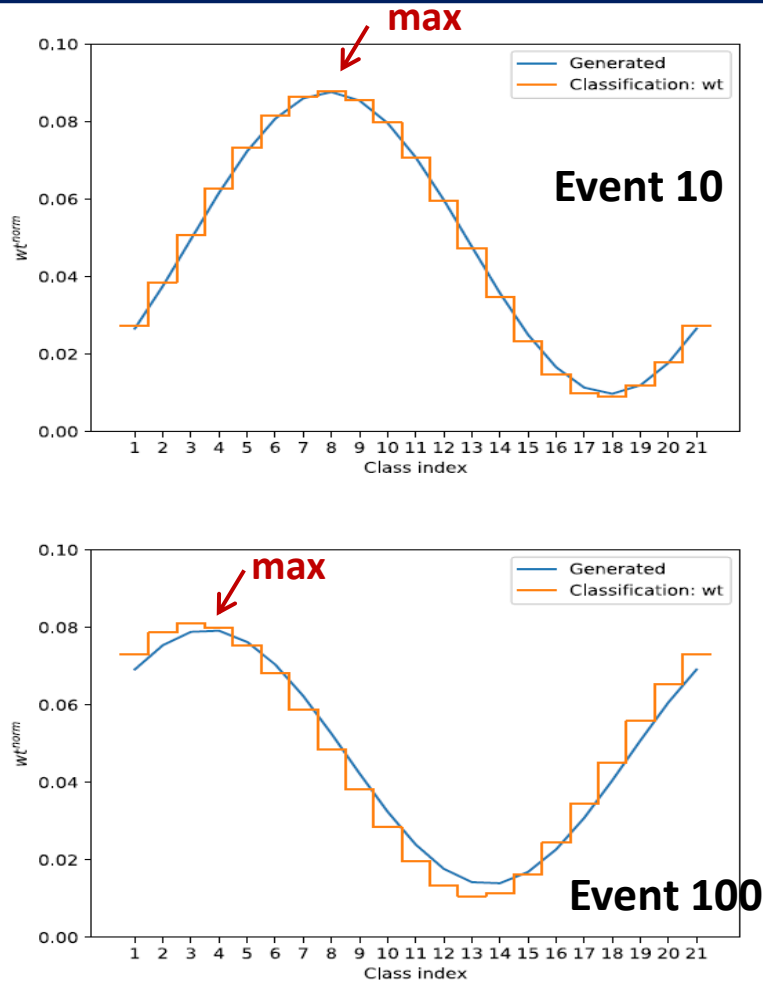


FIG. 5. Normalized to probability spin weight wt^{norm} , predicted (orange steps) and true (blue line), as a function of α_i^{CP} for two example events (top and bottom plots). DNN was trained with $N_{\text{class}} = 21$ spanning range $(0, 2\pi)$.

$$l_2 = \sum_{k=1}^{N_{\text{evt}}} \frac{\sqrt{\int_0^{2\pi} (wt_k^{\text{norm}}(\alpha^{CP}) - p_k(\alpha^{CP}))^2 d\alpha^{CP}}}{N_{\text{evt}}}.$$

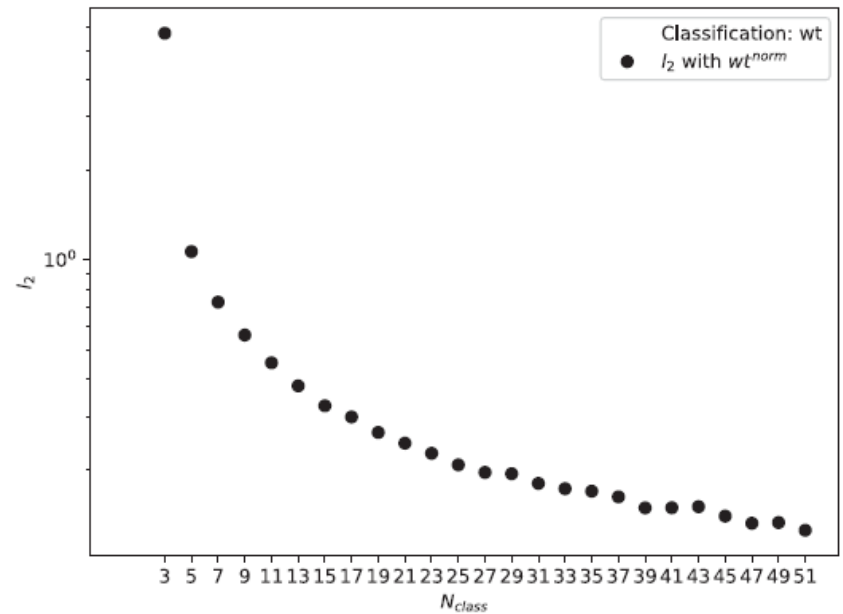


FIG. 6. The l_2 norm, quantifying difference between true and predicted spin weight wt^{norm} , as a function of class multiplicity N_{class} .

Multiclass classification: Learning spin weight

Δ_{class} - difference between most probable predicted class and most probable true class

Periodicity of the functional form taken into account.

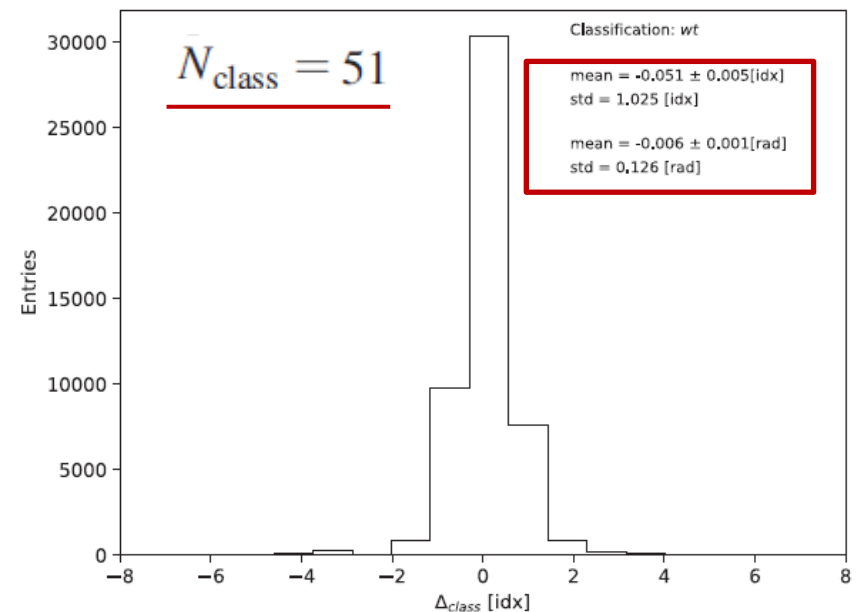
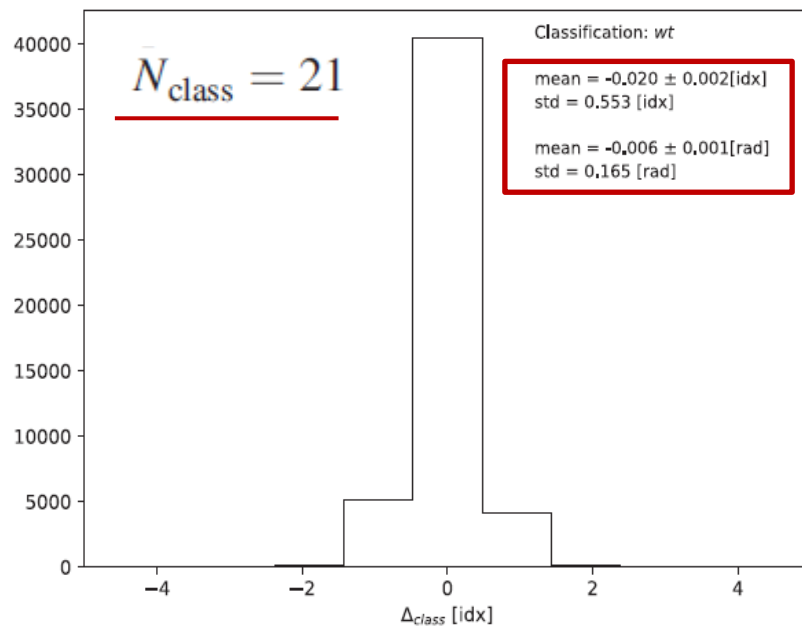
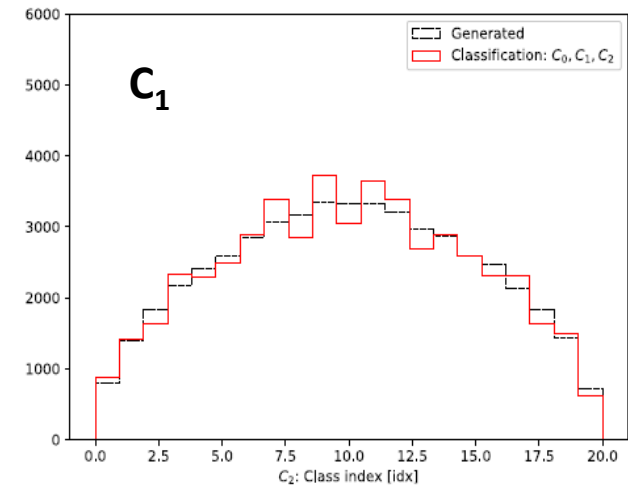
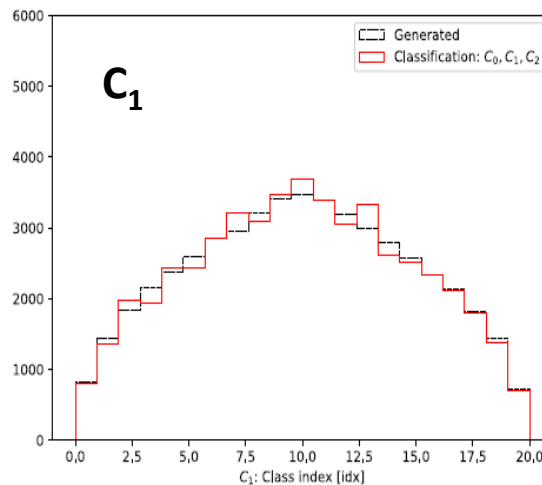
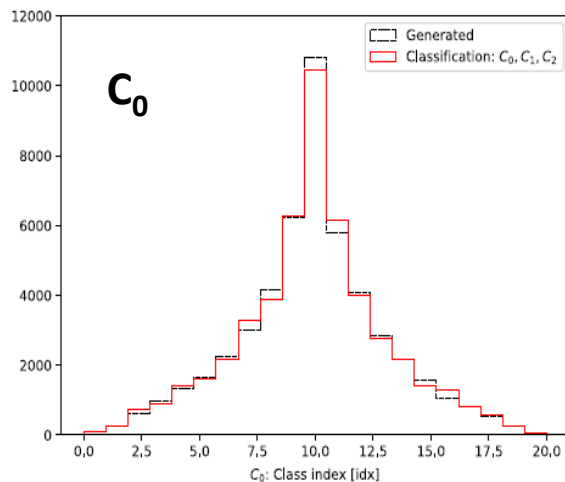
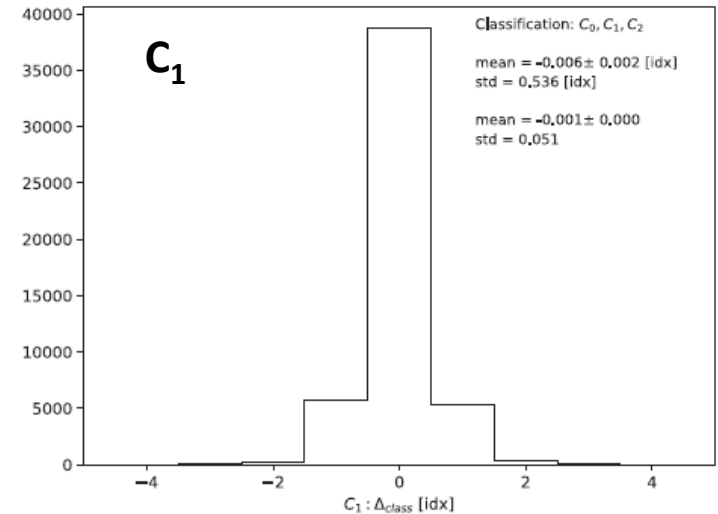


FIG. 7. Distribution of $\Delta_{\text{class}}^{\text{max}}$ between predicted most probable class and true most probable class.

Multiclass classification: learning C_0, C_1, C_2

Δ_{class} - difference between most probable predicted class and most probable true class

$$wt = C_0 + C_1 \cdot \cos(\alpha^{CP}) + C_2 \cdot \sin(\alpha^{CP}),$$

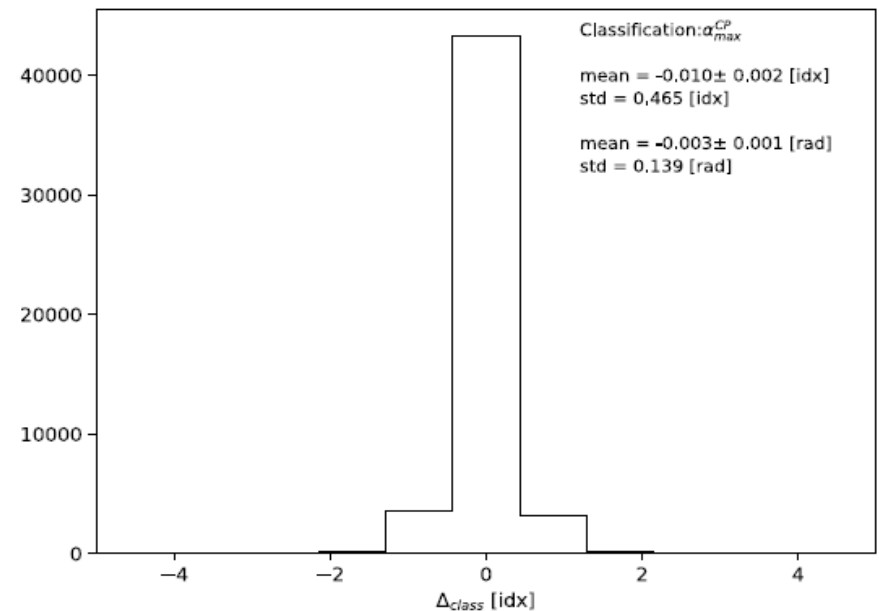
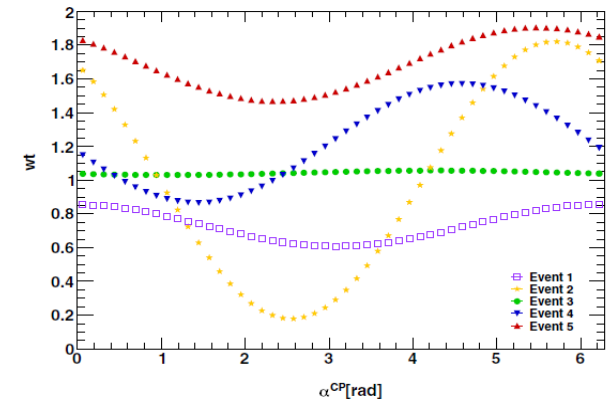
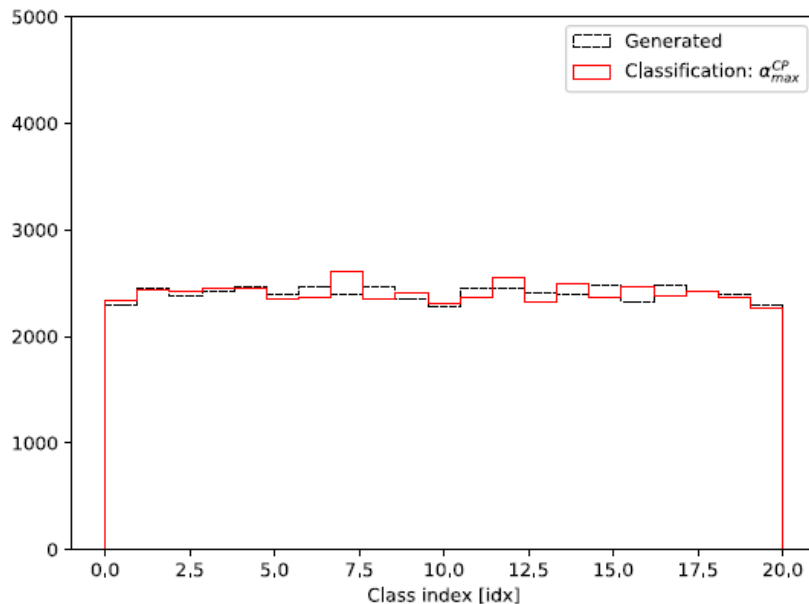


Multi-class classification: learning the $\alpha_{\max}^{\text{CP}}$

Δ_{class} - difference between most probable predicted class and most probable true class

Closure test:

flat distribution, sample is unpolarised

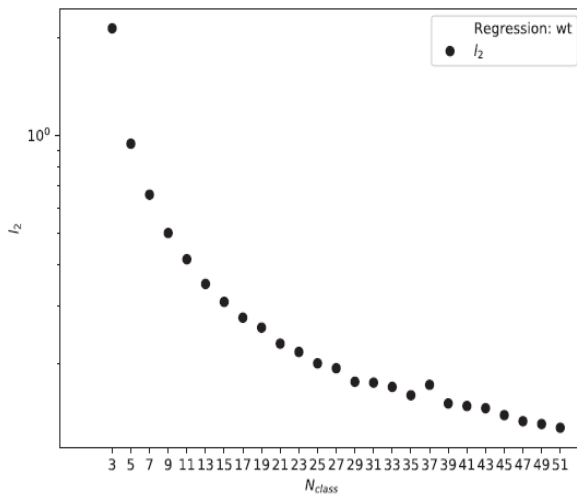


Regression: learning wt, C_i or $\alpha_{\max}^{\text{CP}}$

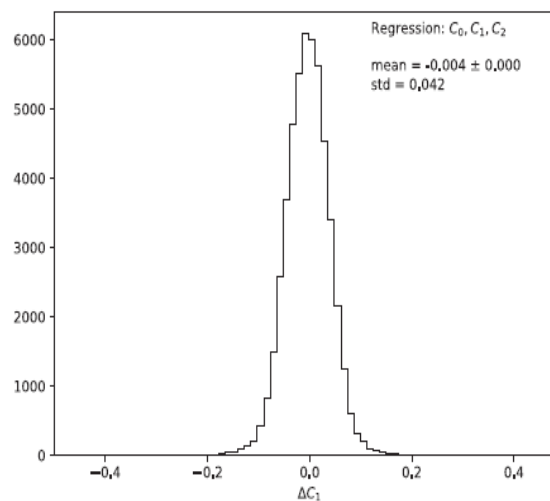
Same tasks formulated as regression problems:

- learning per-event spin weights (discretised)
- learning C_i coefficients (continuous)
- learning most preferred mixing angle (continuous)

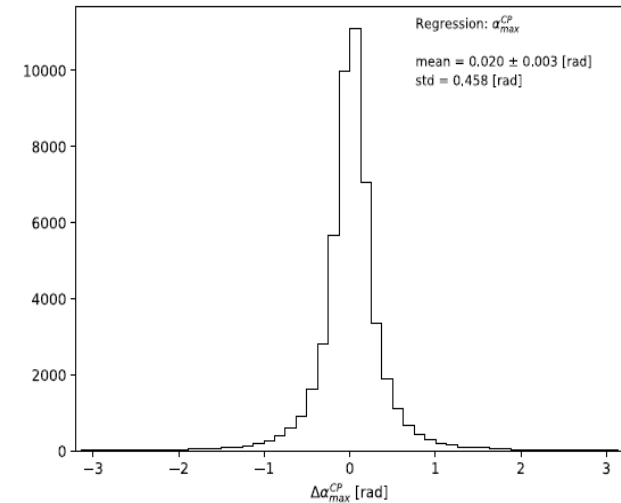
I_2 for spin weight



C_1 coefficient



$\alpha_{\max}^{\text{CP}}$



Regression vs classification: very comparable

Table 3: The mean and standard deviations of ΔC_i , the difference between generated and predicted C_i , obtained from *DNN* with classification and regression methods for $N_{class} = 51$.

Coefficients	Classification	Regression
ΔC_0	mean = 0.000 std = 0.038	mean = 0.004 std = 0.029
ΔC_1	mean = 0.001 std = 0.051	mean = -0.004 std = 0.042
ΔC_2	mean = -0.003 std = 0.051	mean = -0.04 std = 0.042

Table 4: The mean and standard deviation of $\Delta \alpha_{max}^{CP}$, the difference between true and predicted α_{max}^{CP} , obtained from *DNN* with classification and regression methods.

Method	Classification	Regression
Using w_t	mean = -0.006 ± 0.001 [rad] std = 0.126 [rad]	mean = 0.000 ± 0.001 [rad] std = 0.137 [rad]
Using C_0, C_1, C_2	mean = 0.000 ± 0.001 [rad] std = 0.153 [rad]	mean = -0.001 ± 0.001 [rad] std = 0.138 [rad]
Direct	mean = - 0.003 [rad] std = 0.139 [rad]	mean = 0.020 [rad] std = 0.458 [rad]

classification performing better

Summary

DNN was trained to predict:

- Probability that event is of class A given alternative class B
- Spin weight as a function of the CP mixing angle
- Decay configuration dependent coefficients of spin weight formula sensitive to CP mixing angle
- The most preferred mixing angle

Problem was formulated as classification or regression task.

Achieved comparable performance in case complete information on the outgoing particle 4-vectors is provided. For binary classification tried also more realistic feature lists (not including or parametrising neutrino momenta or parametrising tau direction).

Would be very interesting too see some of those ideas tried out in the experimental reality.

Analysis code available from:

- PDR(2019) paper: https://github.com/klasocha/HiggsCP/tree/PRD_2019
- PRD(2021) paper: https://github.com/klasocha/HiggsCP/tree/klasocha_CPmix

DNN technical details

DNN architecture

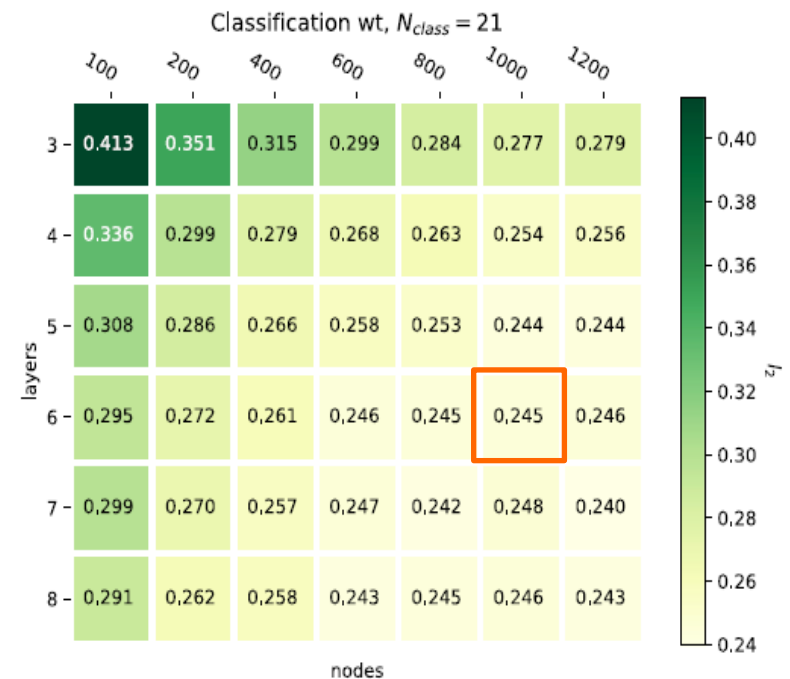
The structure of the event is represented as follows:

$$x_i = (f_{i,1}, \dots, f_{i,D}), \quad w_{a_i}, w_{b_i}, \dots, w_{m_i}$$

The $f_{i,1}, \dots, f_{i,D}$ represent numerical features and $w_{a_i}, w_{b_i}, w_{m_i}$ are weights proportional to the likelihoods that an event comes from a class A, B, \dots, M , each representing different α^{CP} mixing angle. The $\alpha^{CP} = 0, 2\pi$ corresponds to scalar CP state and $\alpha^{CP} = \pi$ to pseudoscalar CP state. Tl

The network architecture consists of 6 hidden layers, 1000 nodes each with ReLU activation functions and is initialized with random weights. Such architecture has been found as a good trade-off between the performance and computation time

procedure is optimized using a variant of stochastic gradient descent algorithm called Adam [32] and batch normalization [33].



DNN: Loss function

Classification:

Cross entropy of valid values and neural network predictions. Common choice in case of binary and multiclass classification models.

$$\text{Loss} = \sum_{k=1}^{N_{\text{evt}}} \sum_{i=1}^{N_{\text{class}}} y_{i,k} \log(p_{i,k}),$$

Regression:

- mean squared error (for discretised wt, and Ci)

$$\text{Loss} = \sum_{k=1}^{N_{\text{evt}}} \sum_{i=1}^{i=N} (y_{i,k} - p_{i,k})^2,$$

- reduce mean for $\alpha^{\text{CP}}_{\text{max}}$

$$\text{Loss} = \sum_{k=1}^{N_{\text{evt}}} (1 - \cos(y_k - p_k)),$$

DNN: training and validation

