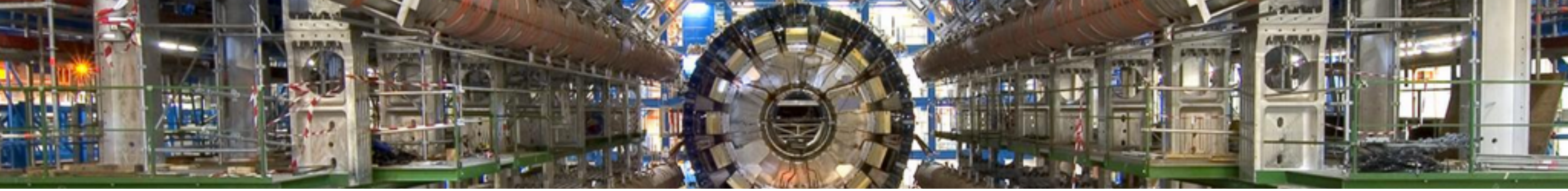
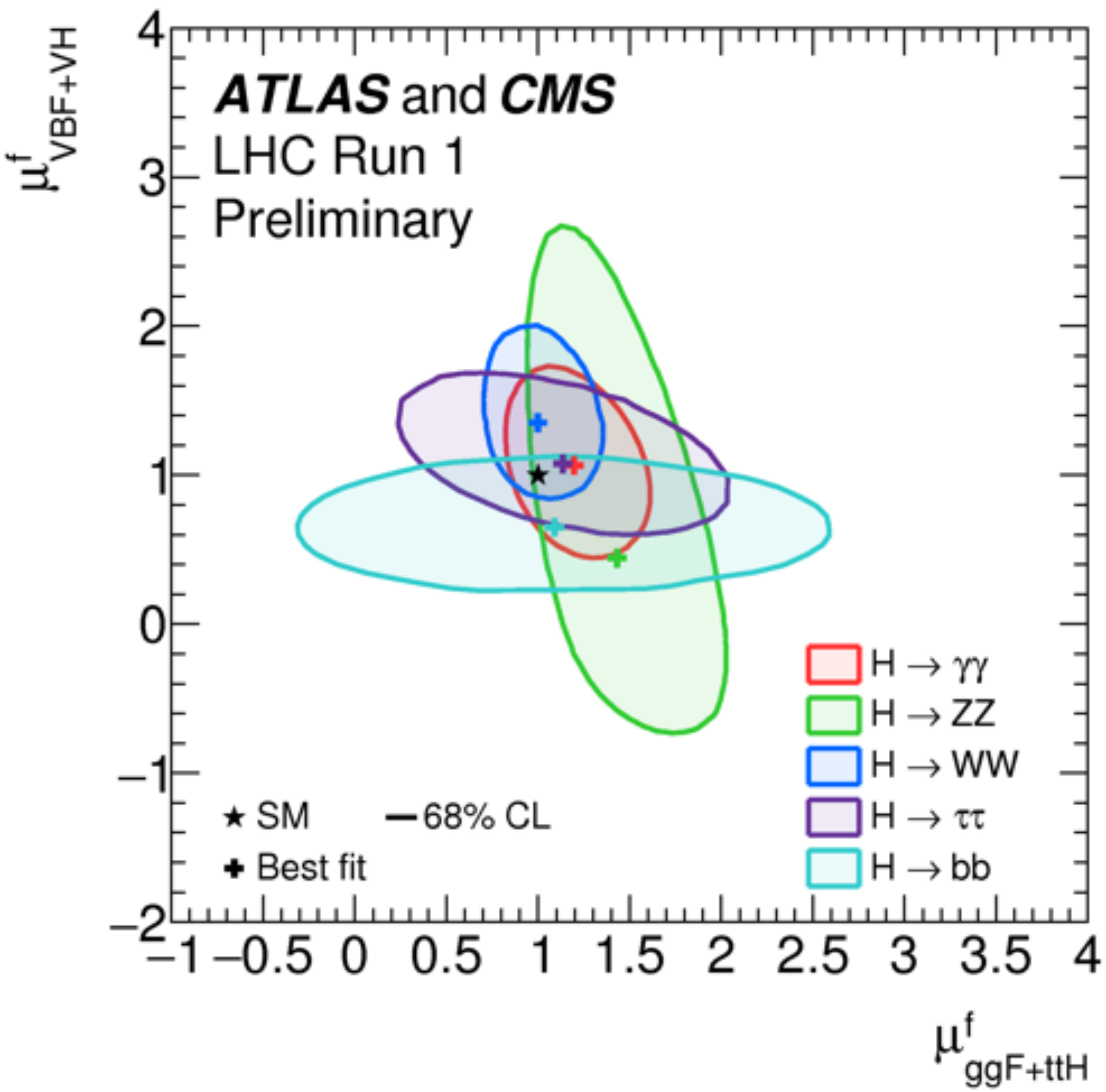


Optimizing Higgs boson CP measurement in $H \longrightarrow \tau\tau$ decay with ML techniques

Brian Le
E. Barberio, E. Richter-Was, Z. Was, D. Zanzi
(supported in part by European Union under the Grant Agreement PITNGA2012316704 (HiggsTools))

THE UNIVERSITY OF MELBOURNE
The Henryk Niewodniczański Institute of Nuclear Physics Polish Academy of Sciences

What Higgs couplings are measured?



Channel	Signal strength [μ]		Signal significance [σ]	
	from results in this paper (Section 5.2)			
	ATLAS	CMS	ATLAS	CMS
$H \to \gamma\gamma$	$1.15^{+0.27}_{-0.25}$	$1.12^{+0.25}_{-0.23}$	5.0	5.6
	$(^{+0.26}_{-0.24})$	$(^{+0.24}_{-0.22})$	(4.6)	(5.1)
$H \to ZZ \to 4\ell$	$1.51^{+0.39}_{-0.34}$	$1.05^{+0.32}_{-0.27}$	6.6	7.0
	$(^{+0.33}_{-0.27})$	$(^{+0.31}_{-0.26})$	(5.5)	(6.8)
$H \to WW$	$1.23^{+0.23}_{-0.21}$	$0.91^{+0.24}_{-0.21}$	6.8	4.8
	$(^{+0.21}_{-0.20})$	$(^{+0.23}_{-0.20})$	(5.8)	(5.6)
$H \to \tau\tau$	$1.41^{+0.40}_{-0.35}$	$0.89^{+0.31}_{-0.28}$	4.4	3.4
	$(^{+0.37}_{-0.33})$	$(^{+0.31}_{-0.29})$	(3.3)	(3.7)
$H \to bb$	$0.62^{+0.37}_{-0.36}$	$0.81^{+0.45}_{-0.42}$	1.7	2.0
	$(^{+0.39}_{-0.37})$	$(^{+0.45}_{-0.43})$	(2.7)	(2.5)
$H \to \mu\mu$	-0.7 ± 3.6	0.8 ± 3.5		
	(± 3.6)	(± 3.5)		
ttH production	$1.9^{+0.8}_{-0.7}$	$2.9^{+1.0}_{-0.9}$	2.7	3.6
	$(^{+0.72}_{-0.66})$	$(^{+0.88}_{-0.80})$	(1.6)	(1.3)

arxiv:1606.02266

For **H→$\tau\tau$**, we aren't quite at discovery level yet

Coupling and Spin-CP

H⟶VV channels have claimed discovery and made measurements of coupling and CP properties - still no direct measurements to fermions

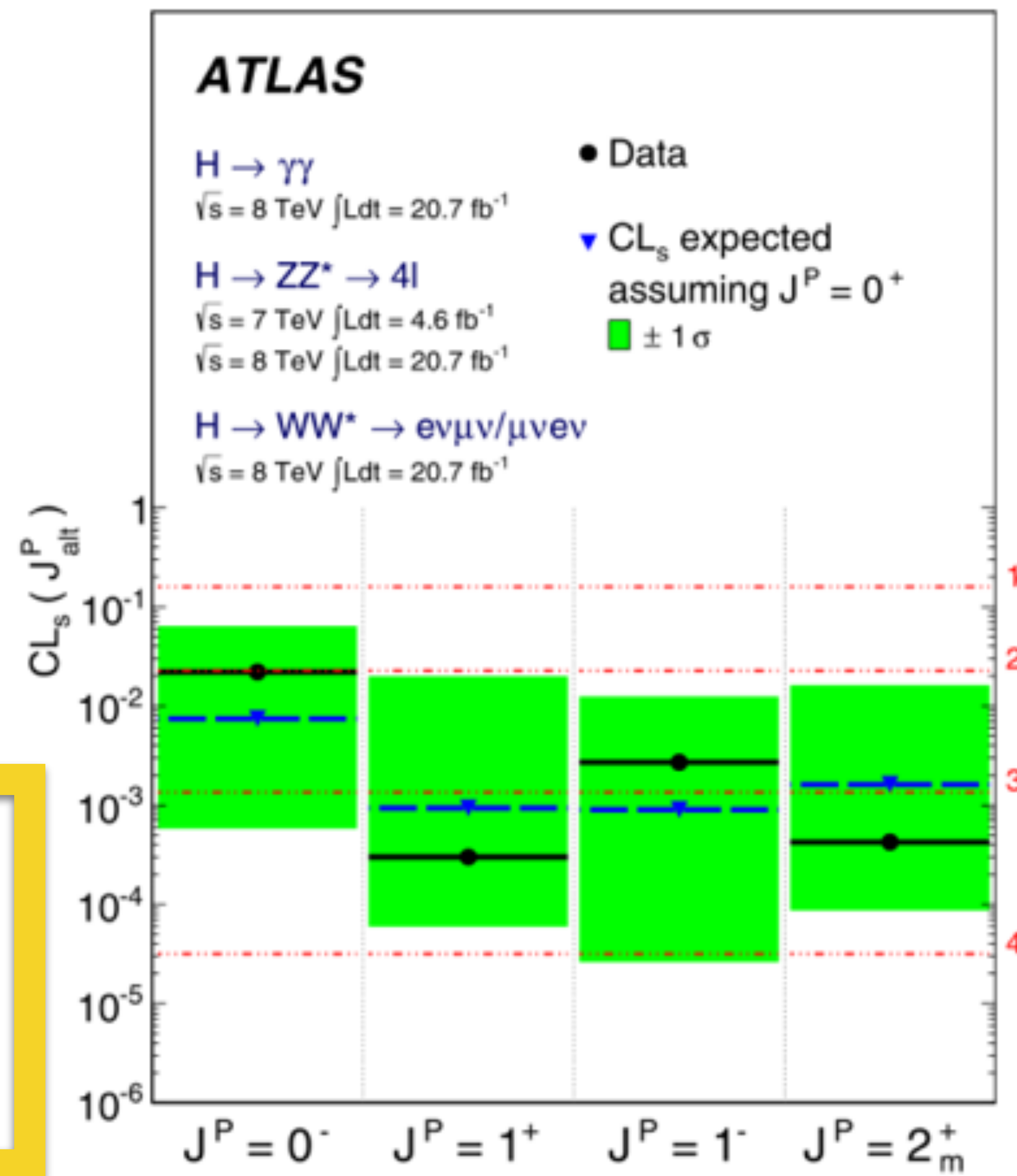
H⟶$\tau\tau$ will be the first to measure the coupling and CP properties in fermionic decays

- H⟶VV couplings not sensitive to tree-level couplings - only can exclude pure states
- **H⟶$\tau\tau$** is sensitive to tree level couplings to **CP-odd Higgs boson - sensitive to CP mixing**

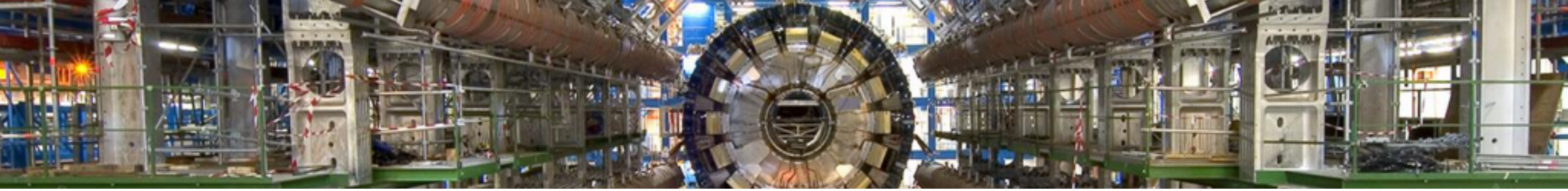
$$\mathcal{L}_Y^{tau} = -g_\tau \left(\cos\phi_\tau \cdot \bar{\tau}\tau + \sin\phi_\tau \cdot \bar{\tau} i \gamma_5 \tau\right) h$$

$$\Gamma(h_{mix} \to \tau^+\tau^-) \sim 1 - s_\parallel^{\tau^+} s_\parallel^{\tau^-} + s_\perp^{\tau^+} R(2\phi_\tau)\, s_\perp^{\tau^-}$$

where R is a rotation in the x-y plane



arXiv:1307.1432



A little Theory (and History!)

Now to the blackboard…

References

- Potential for optimizing Higgs boson CP measurement in H to tau tau decay at LHC and ML techniques (R. Józefowicz, E. Richter-Was, Z. Was) Phys. Rev. D 94 (2016): arxiv:1608.02609 ← Cutting edge tools of Google used (and cross-checked)
- Probing the CP nature of the Higgs boson at linear colliders with τ spin correlations; the case of mixed scalar–pseudoscalar couplings (K. Desch, A. Imhof, Z. Was, M. Worek) Eur. Phys. J. C29 (2003): arxiv:0307331
- Measuring Higgs parity with τ⟶ρν decays (G. R. Bower, T. Pierzchala, Z. Was, M. Worek) Phys. Lett. B543 (2002) : arxiv:0204292

- Extra references:
- Prospects of constraining the Higgs CP nature in the tau decay channel at the LHC (S. Berge, W. Bernreuther, S. Kirchner) Phys. Rev. D92 (2015) : arxiv:1510.03850
- Reconstruction of hadronic decay products of tau leptons with the ATLAS experiment (ATLAS collaboration) Eur. Phys. J C 76(5) (2016): arxiv:1512.05955
- Tensorflow (low-level neural network training software): https://www.tensorflow.org/
- Keras (high-level interface for Tensorflow and Theano): https://keras.io/

Tau Branching Ratios

Our focus

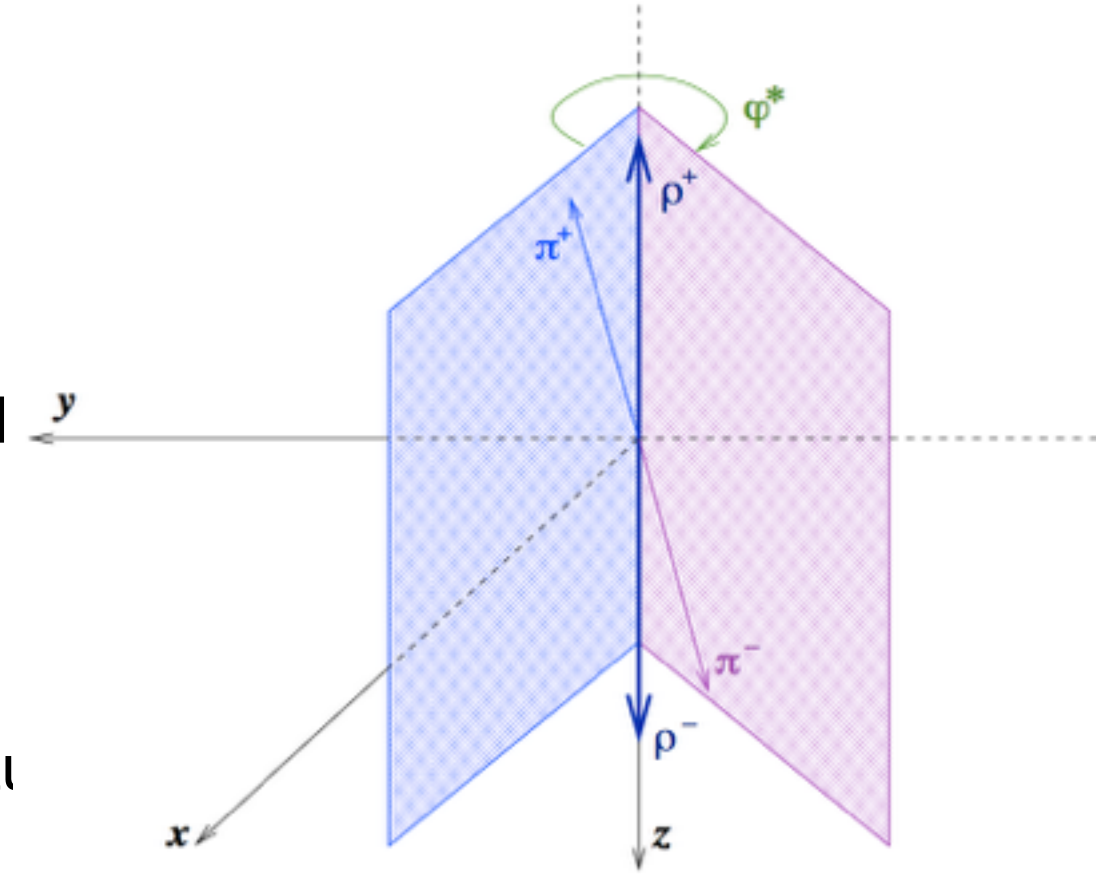
- Feasible
- Maybe Useful

category	decay mode	$\mathcal{B}$ [%]	nomenclature
hadronic 1-prong	$\tau^- \to \pi^- \nu_\tau$	**10.8**	1p0n
	$\tau^- \to K^- \nu_\tau$	0.7	1p0n
	$\tau^- \to \rho^-(\to \pi^- \pi^0)\nu_\tau$	**25.5**	1p1n
	$\tau^- \to K^{-\star}(\to K^- \pi^0)\nu_\tau$	0.4	1p1n
	$\tau^- \to a_1^-(\to \pi^- \pi^0 \pi^0)\nu_\tau$	**9.3**	1pXn
	$\tau^- \to K^- \pi^0 \pi^0 \nu_\tau$	0.1	1pXn
	$\tau^- \to \pi^- \pi^0 \pi^0 \pi^0 \nu_\tau$	1.1	1pXn
	$\tau^- \to hK^0_S \geq 0$ neutrals ν_τ	0.9	1pXk
hadronic 3-prong	$\tau^- \to \pi^- \pi^- \pi^+ \nu_\tau$ (mostly via a_1^-)	**9.3**	3p0n
	$\tau^- \to \pi^- \pi^- \pi^+ \pi^0 \nu_\tau$	**4.6**	3pXn
	$\tau^- \to hhhK^0 \nu_\tau$	0.2	3pXk
hadronic ≥5-prong	$\tau^- \to \geq 5h \geq 0$ neutrals ν_τ	0.1	(none)
leptonic	$\tau^- \to e^- \nu_\tau \nu_e$	17.8	(none)
	$\tau^- \to \mu^- \nu_\tau \nu_\mu$	17.4	(none)

From 6% to 12% of $H \to \tau\tau$

Observable for ρ decays

Method :
Use **secondary decay products** component to **form the decay plane**



Pro : Independent from impact parameter method

Con : So far prepared for **single decay mode** (6.5% of all H⟶$\tau\tau$ decays)

Provides strongest observable for ρ decays of di-tau system. Want to extend method to decays with intermediate resonance (three charged π final state)

arxiv:0307331

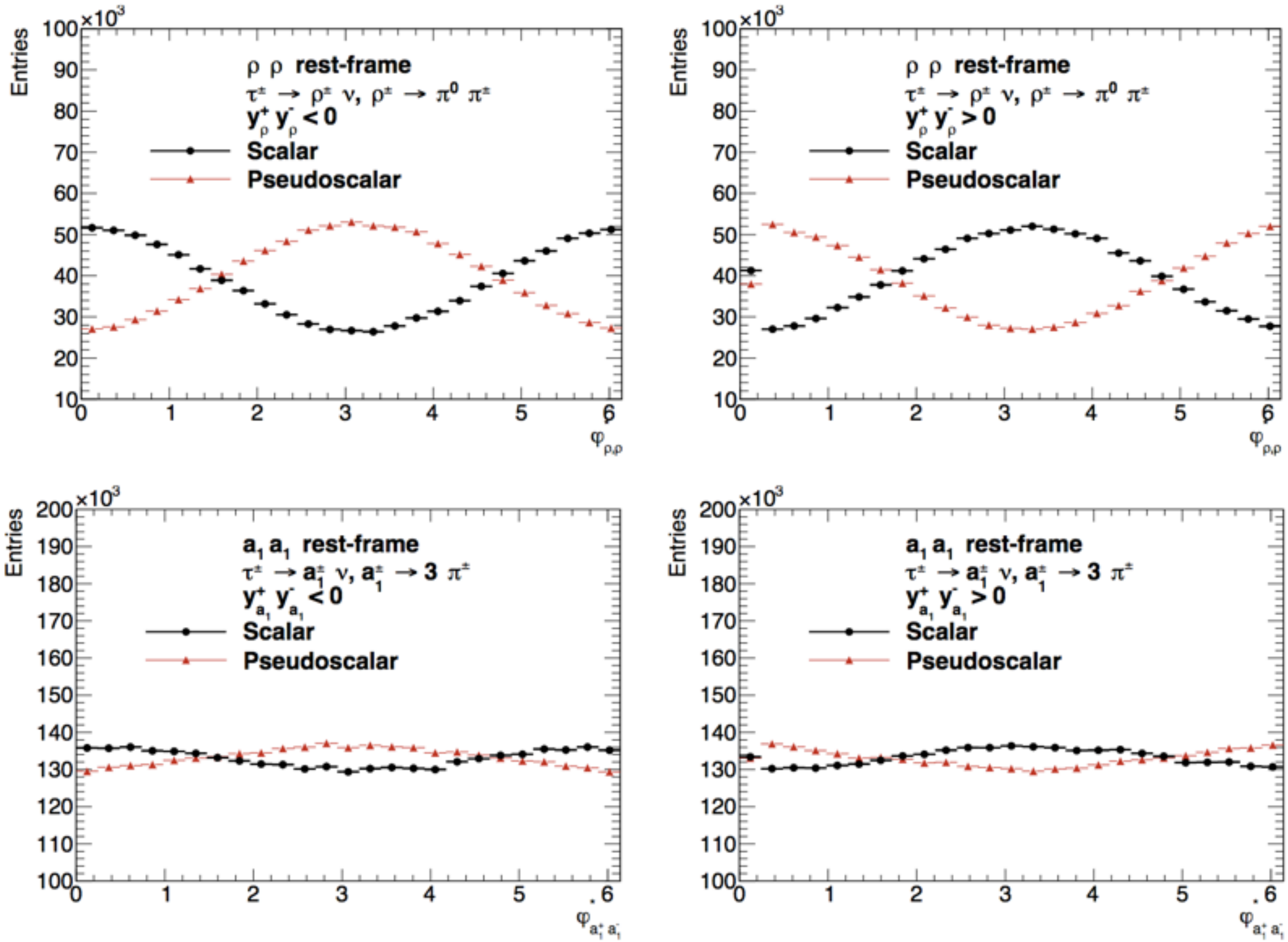
$$\tau^{\pm} \longrightarrow a_1 \, \nu$$
$$\longrightarrow \rho \, \pi^{\pm} \nu$$
$$\longrightarrow \pi^{\pm}\pi^{\mp}\pi^{\pm} \nu$$

New: Substitute the neutral π in the above method with the neutral ρ, it further decays into $\pi^{\pm}\pi^{\mp}$ so another plane can be defined.

Now to the blackboard…

ρ-ρ vs a_1-a_1

(for a_1 small amplitude but many distributions)



Deep Learning NN

R. Józefowicz (Google (NY), now at Open AI (SFO)) developed neural network model with Tensorflow (Google project for various non-HEP applications).

Z. Was and E. Richter-Was found promising separation utilising the neural network between scalar and pseudoscalar.

Why a neural network?

- Problem is very multidimensional (a_1-a_1 can have 16 possible acoplanar angles and 8 y variables)
- Separation amplitudes are small for each individual acoplanar angle
- NN allows for non-linear connections between all variables

Input samples of Pythia generated $H \rightarrow \tau\tau$ (τ decays simulated with TAUOLA) and weights for scalar and pseudoscalar angles generated with TauSpinner.
Systematic errors need to be studied.

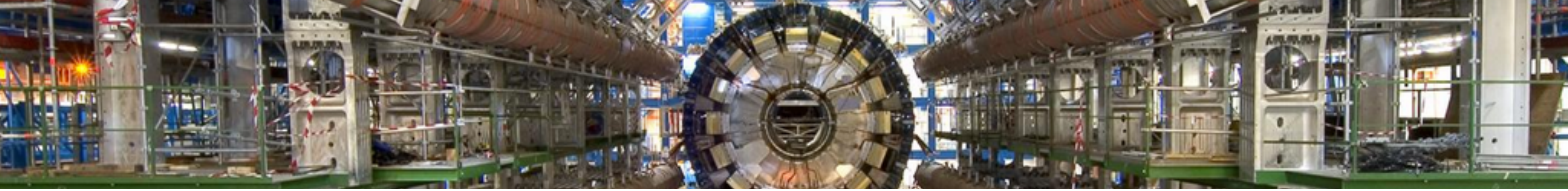
Results from 1608.02609

Features/variables	Decay mode: $\rho^{\pm}-\rho^{\mp}$ $\rho^{\pm}\to\pi^{0}\,\pi^{\pm}$	Decay mode: $a_1^{\pm}-\rho^{\mp}$ $a_1^{\pm}\to\rho^{0}\pi^{\mp}$, $\rho^{0}\to\pi^{+}\pi^{-}$ $\rho^{\mp}\to\pi^{0}\,\pi^{\mp}$	Decay mode: $a_1^{\pm}-a_1^{\mp}$ $a_1^{\pm}\to\rho^{0}\pi^{\pm}$, $\rho^{0}\to\pi^{+}\pi^{-}$
Complete kinematics + ME weights	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

Results show a fair amount of separation. For a_1-a_1 decay mode, it is weaker but given the increase in acceptance by its inclusion, it is worth further investigation.

Seemingly most important class of input are the 4-vectors. Would indicate the neural network can “learn” important features such as y and mass within the model.

Note: All features were boosted to the rest frame of the visible decay products and momenta of the primary resonances **had to be** aligned along the z axis.



Experimental Challenges

Neutrinos

Neutrinos have not provided sensitivity for this measurement (yet). However looking at the polarimetric vector for $\tau \longrightarrow \pi^{\pm}\pi^{0}\,\nu$ decays:

$$h^i = \mathcal{N}\left(2(q \cdot N)q^i - q^2 N^i\right)$$

the second term can not be used directly as neutrinos escape detection. Constraints from missing momentum/energy can be used instead(?)

One option is to use approximations to reconstruct the neutrino momentum (like the collinear approximation). Leads to quartic equations (which are very messy).

Another is to use the neutrinos reconstructed from an estimator for the tautau system, such as the MMC algorithm arxiv:1012.4686 (but this can have biases)

Impact parameter method can be used to constrain the neutrino.

My idea is to let the neural network "learn" what it needs to from the missing transverse energy. For 3π decay channel the formula for h^i is longer.

Latest results: Huge Improvement

Partial use of neutrinos via E_T^{miss}

	No $p_T(H)$ cut			$p_T(H) > 100$ GeV		
	ρ-ρ	ρ-a_1	a_1-a_1	ρ-ρ	ρ-a_1	a_1-a_1
φ^*, y	0.626	0.573	0.539	0.613	0.552	0.525
φ^*, y, m	0.626	0.582	0.557	0.612	0.567	0.539
φ^*, 4-vec	0.638	0.595	0.572	0.627	0.575	0.550
φ^*, y, E_T^{miss}	0.639	0.591	0.537	0.628	0.535	0.522
φ^*, y, m, E_T^{miss}	0.640	0.608	0.592	0.625	0.589	0.535
φ^*, 4-vec, E_T^{miss}	0.712	0.683	0.661	0.684	0.660	0.611

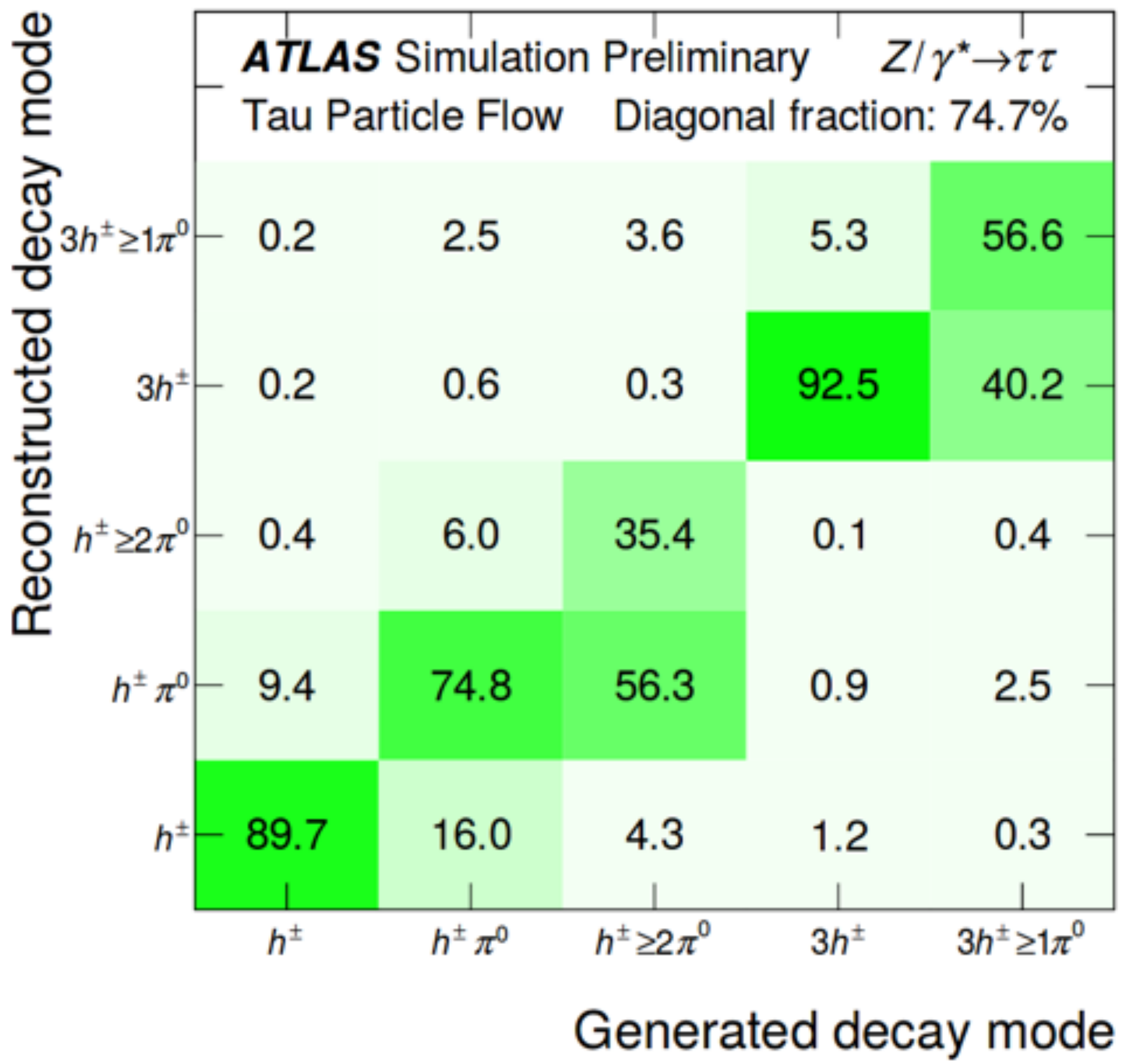
Results depend on model assumptions. Models inspired with results. Fitting setup ⟶ biases. Our algorithms are far less elaborate than human eye/ brain. Problems known since times of Giuseppe Arcimboldo (1572 - 1593) at least.

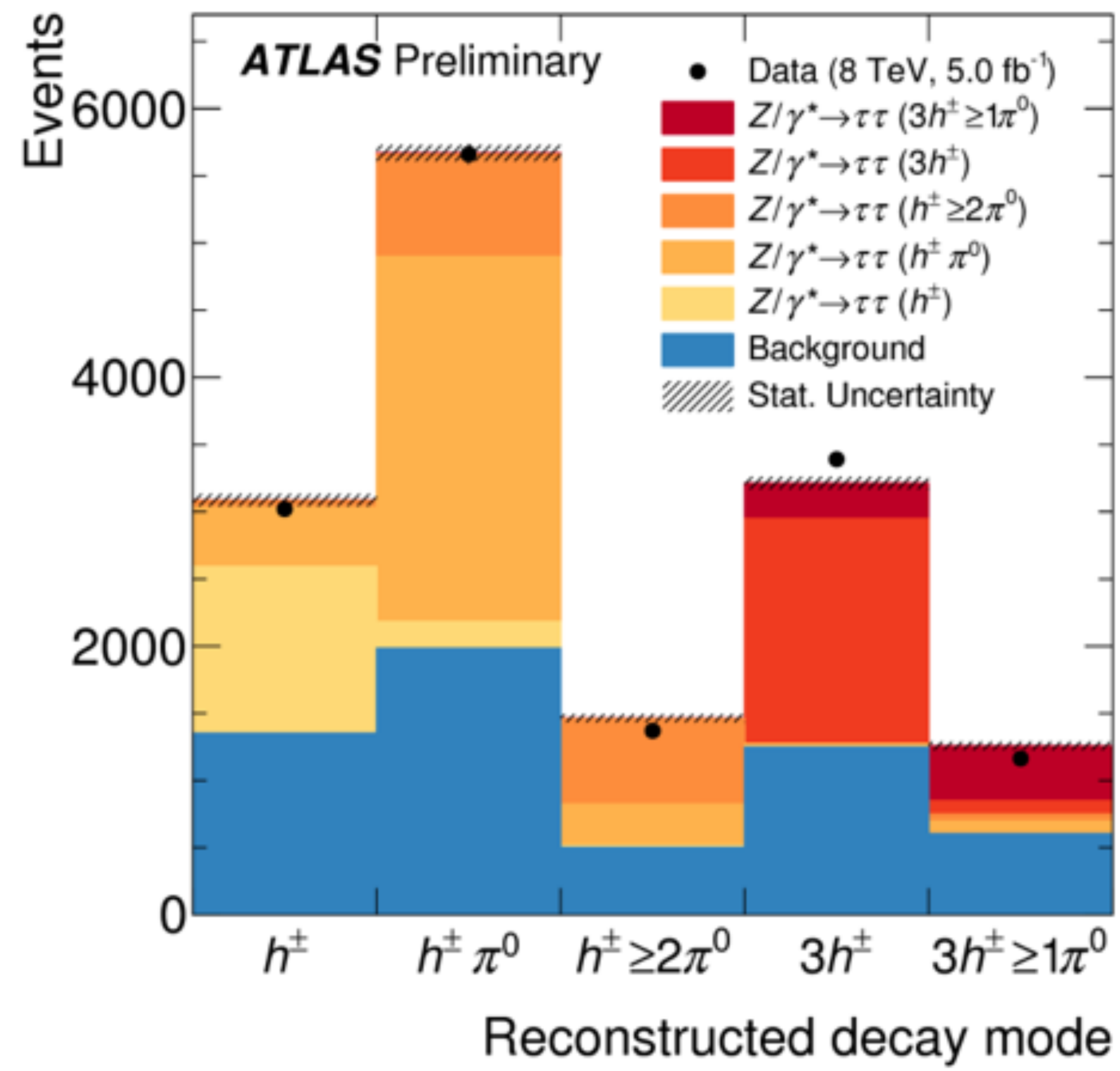
NN brings no improvement for this…

Artificial Neural Networks have spurred remarkable recent progress in image classification and speech recognition. But even though these are very useful tools based on well-known mathematical methods, **we actually understand surprisingly little of why certain models work and others don't**.

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Substructure Reconstruction

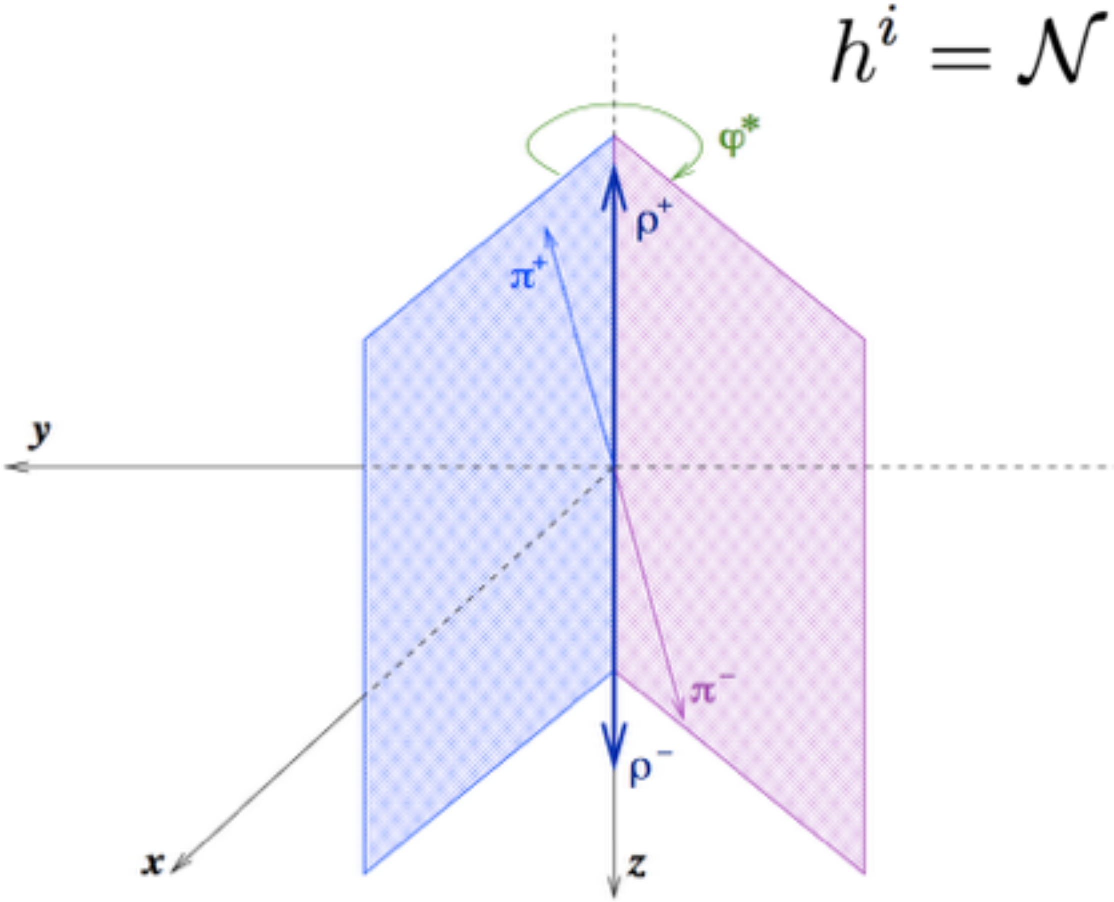




ATL-COM-PHYS-2015-214

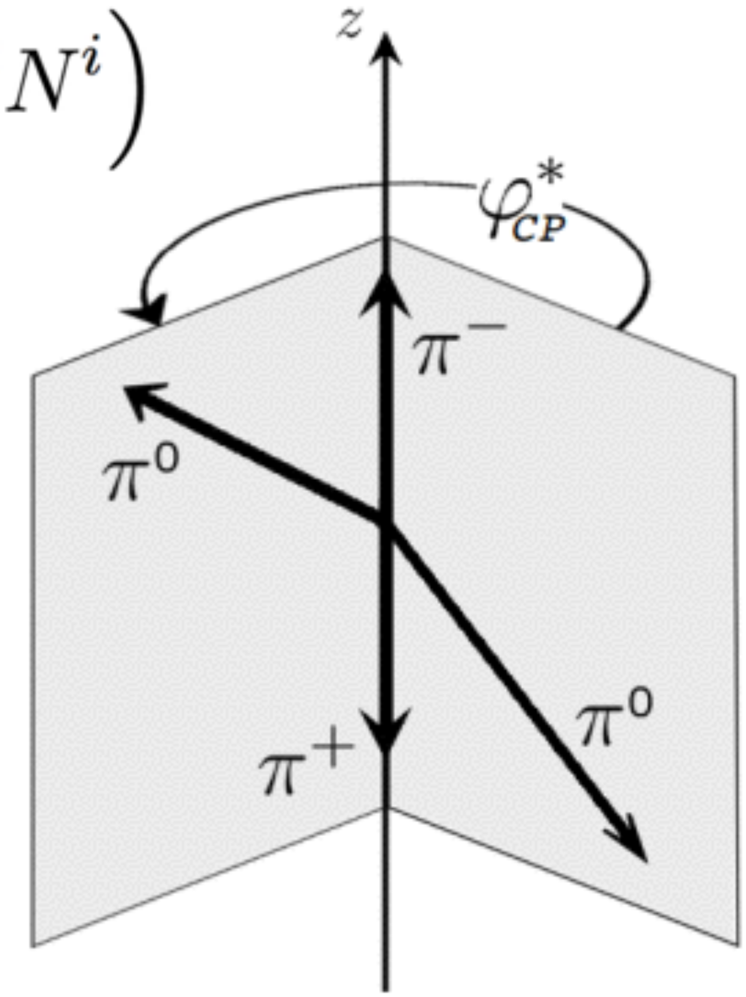
Quick word on the frame

$$h^i = \mathcal{N}\left(2(q \cdot N)q^i - q^2 N^i\right)$$



arxiv:0307331

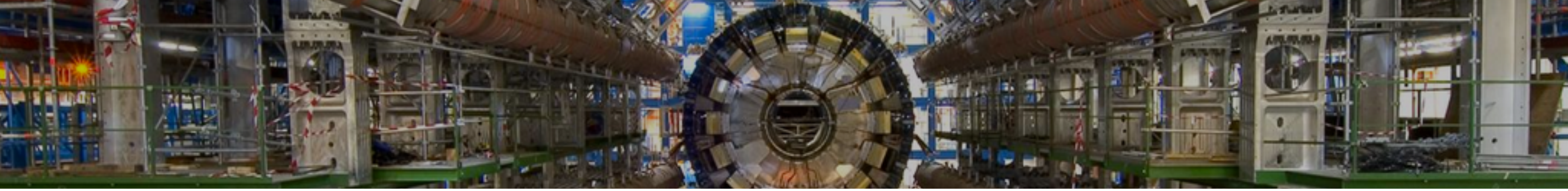
VS



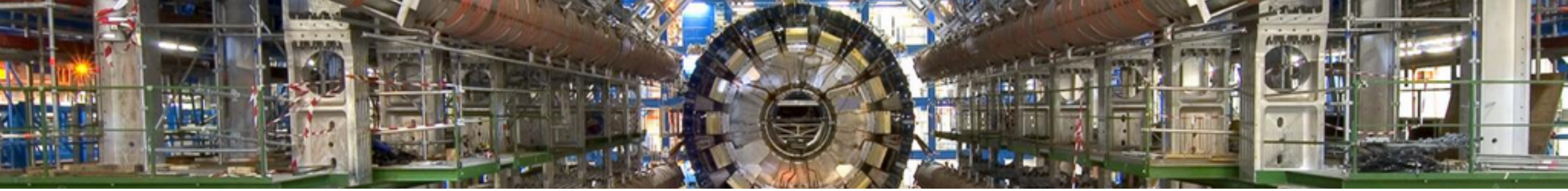
arxiv:1510.03850

Note: from theoretical perspective defining acoplanarities in frame of decay visible decay product or charged pion's frame is not of importance. Two cases may however suffer from distinct experimental systematics

Now to the blackboard…



Backup

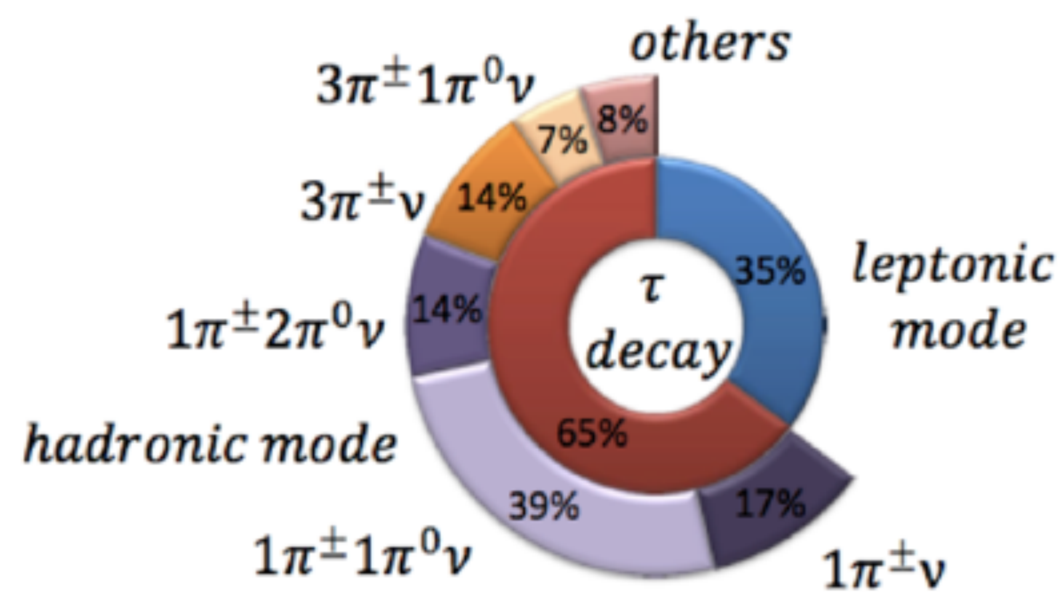


Part 1:

(ATLAS) Tau Reconstruction

Tau Decays (ATLAS)

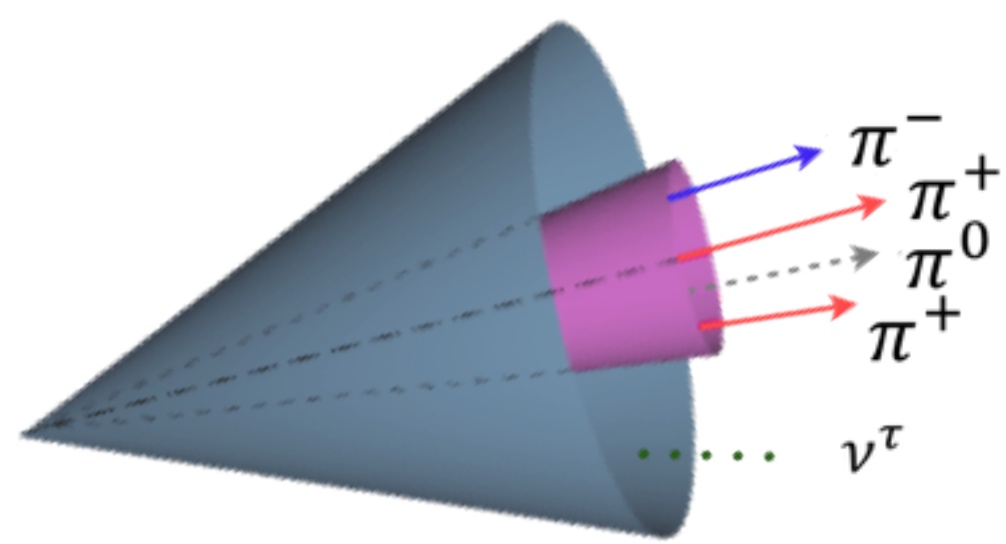
Tau leptons are not reconstructed directly by ATLAS, only the decay products (lepton or hadrons) are detected



Decays to leptons are indistinguishable to prompt leptons

Tau-jet or τ_{had} candidates typically are:

- Highly collimated "jet"
- Odd number of charged tracks ("prongs") with neutral pions



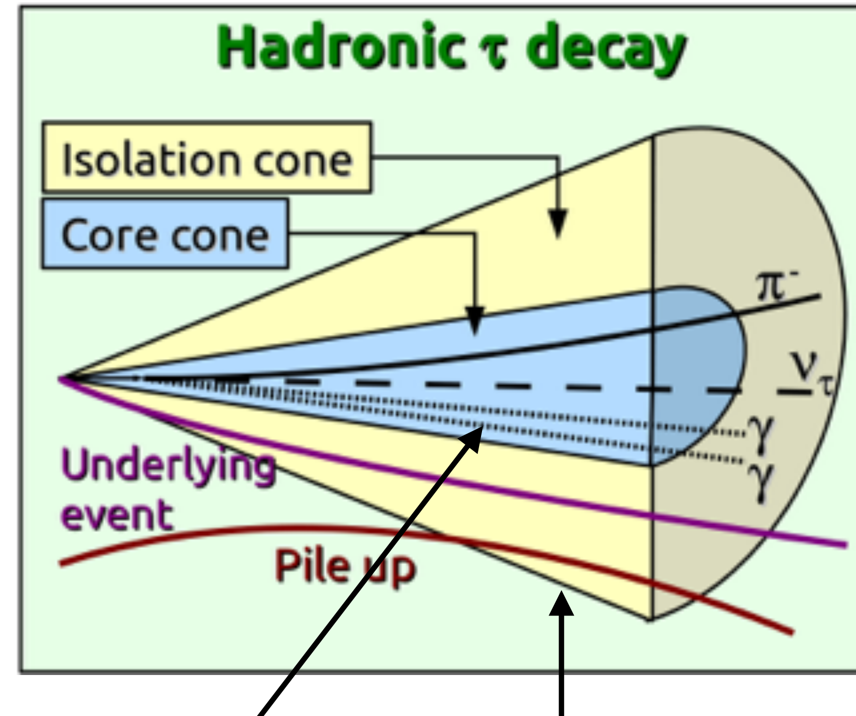
Reconstructed from jet candidates via anti-kT algorithm

Identification via BDT focussed on distinguishing vs QCD jets

Tau Reco + ID (ATLAS)

Tau-jets are reconstructed from jets reconstructed from anti-k_T jets with $\Delta R=0.4$. Tracks are required to be contained within the core cone of $\Delta R<0.2$.

Identification is performed through a multivariate classifier. Three working points (loose, medium, tight) are defined for specific signal efficiencies.

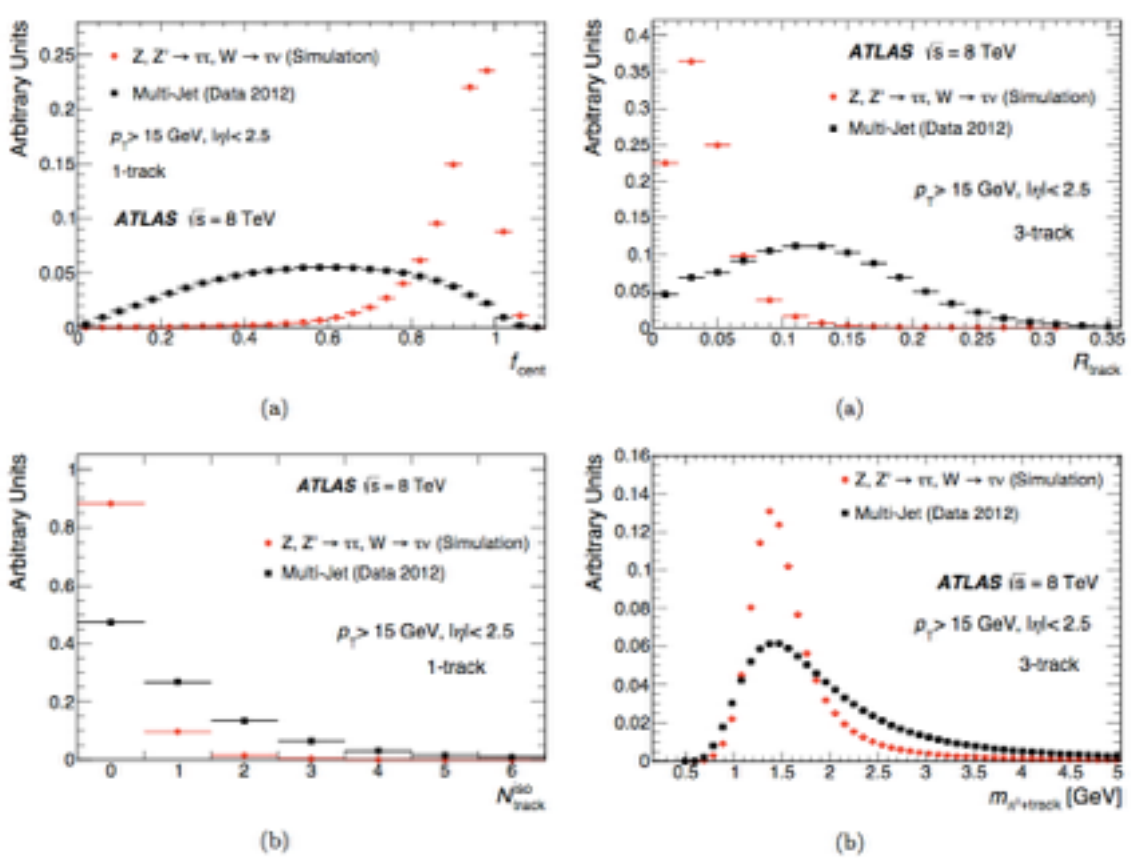


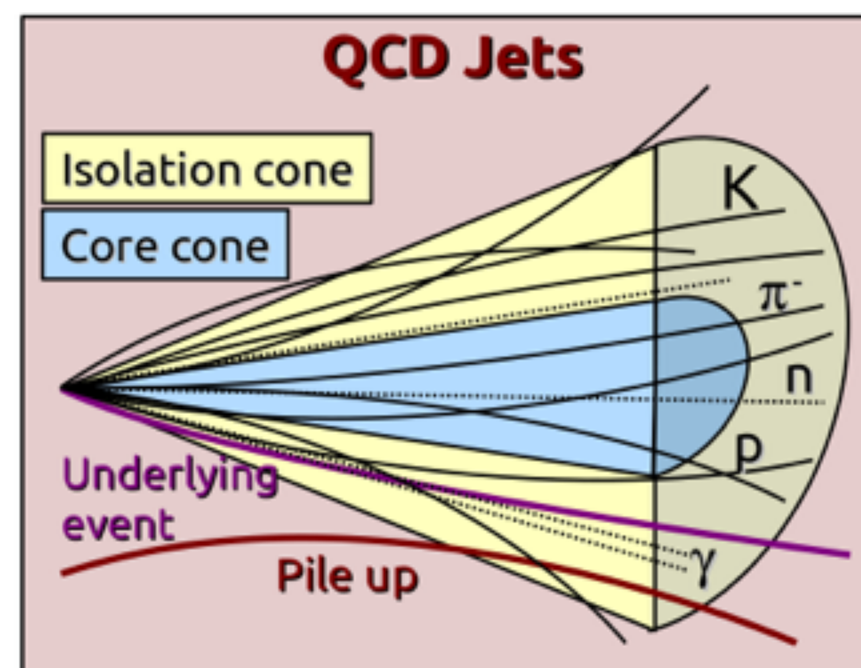
ΔR<0.2 0.2<ΔR<0.4

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$$

Variable	Offline 1-track	Offline 3-track
f_{cent}	●	●
$f^{-1}_{\mathrm{leadtrack}}$	●	●
$R^{0.2}_{\mathrm{track}}$	●	●
$\lvert S_{\mathrm{leadtrack}}\rvert$	●	
$f^{\mathrm{track}}_{\mathrm{iso}}$	●	
ΔR_{Max}		●
$S^{\mathrm{flight}}_{\mathrm{T}}$		●
m_{track}		●
$f^{\mathrm{track-HAD}}_{\mathrm{EM}}$	●	●
$f^{\mathrm{EM}}_{\mathrm{track}}$	●	●
$m_{\mathrm{EM+track}}$	●	●
$p^{\mathrm{EM+track}}_{\mathrm{T}}/p_{\mathrm{T}}$	●	●

[arxiv:1412.7086](https://arxiv.org/abs/1412.7086)





Substructure Reconstruction (ATLAS)

One key development for Run II is to allow for the tau substructure to be reconstructed.

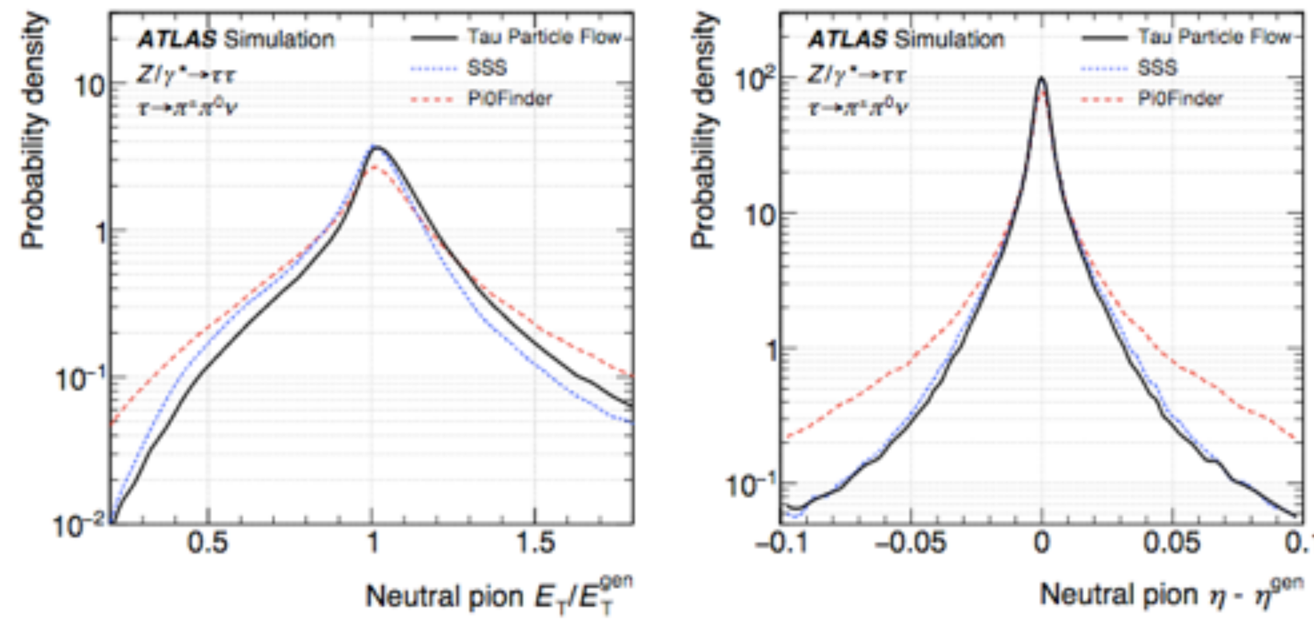
A particle flow approach is taken (rather than using only calo information). Charged hadrons reconstructed using track information, neutrals from calo deposits.

Leads to better four momenta-resolution and allows for classification of tau decay.

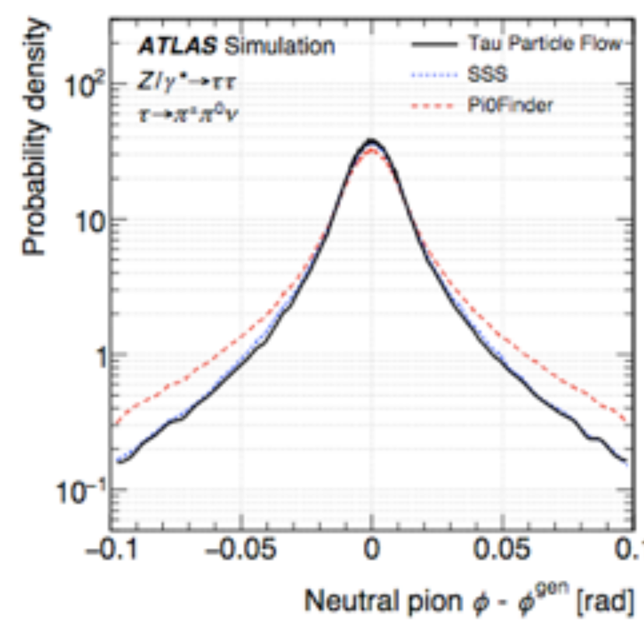
Three BDTs are formed to separate decay modes:

- 1p0n from 1p1n
- 1p1n from 1pXn (so far the most difficult)
- 3p0n from 3pXn

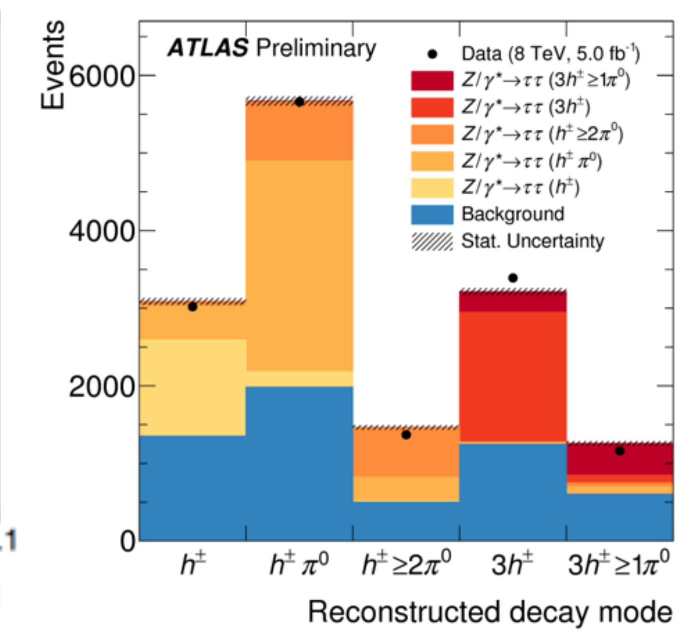
A five way classification is defined and will be critical in forming the structure of the CP measurement.



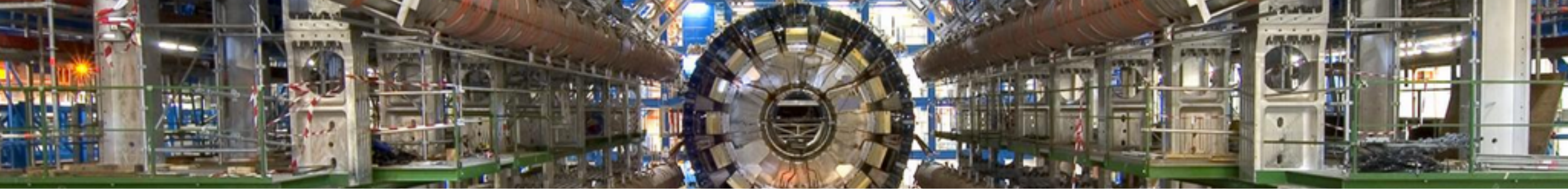
(a) (b)



(c)



ATL-COM-PHYS-2015-214



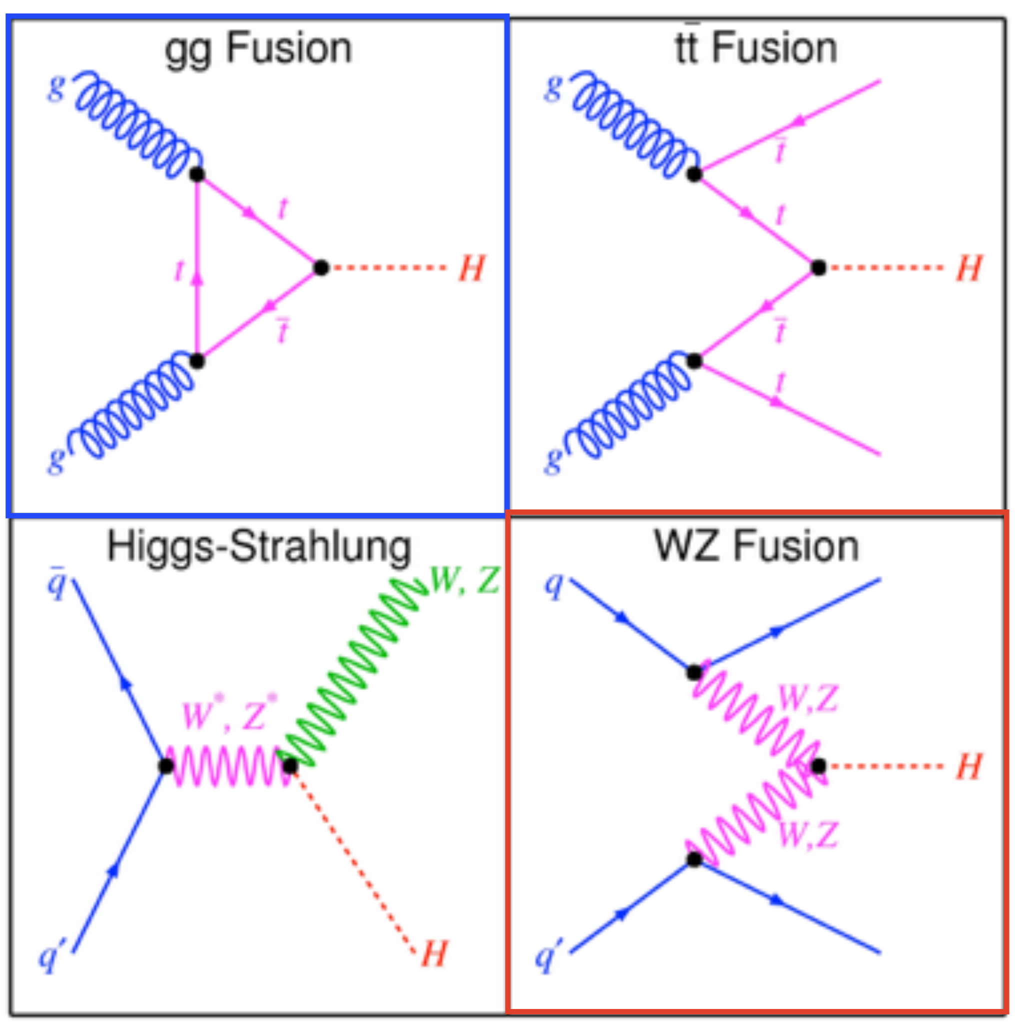
Part 2:

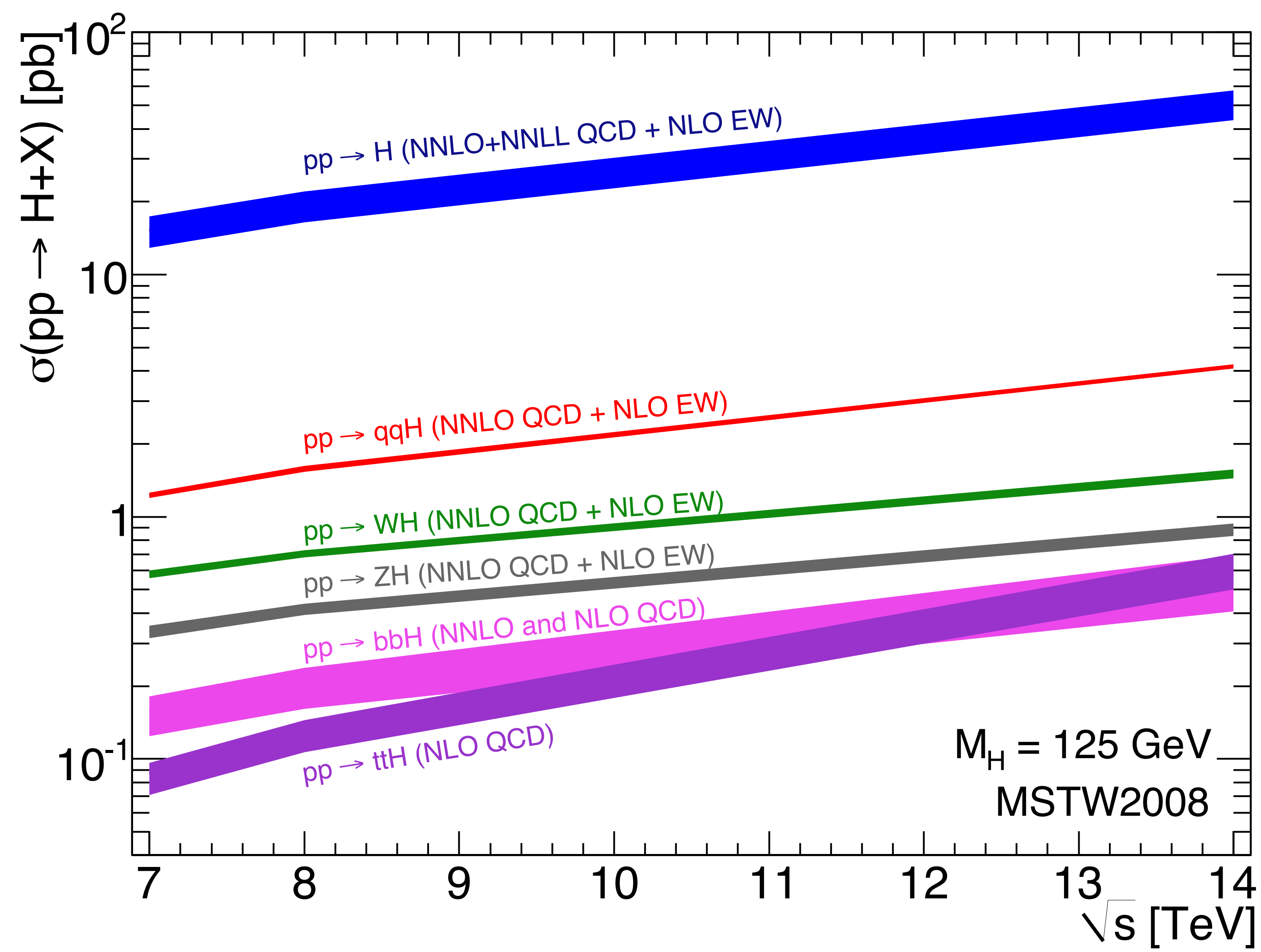
SM Coupling Analysis

H→ττ Signal

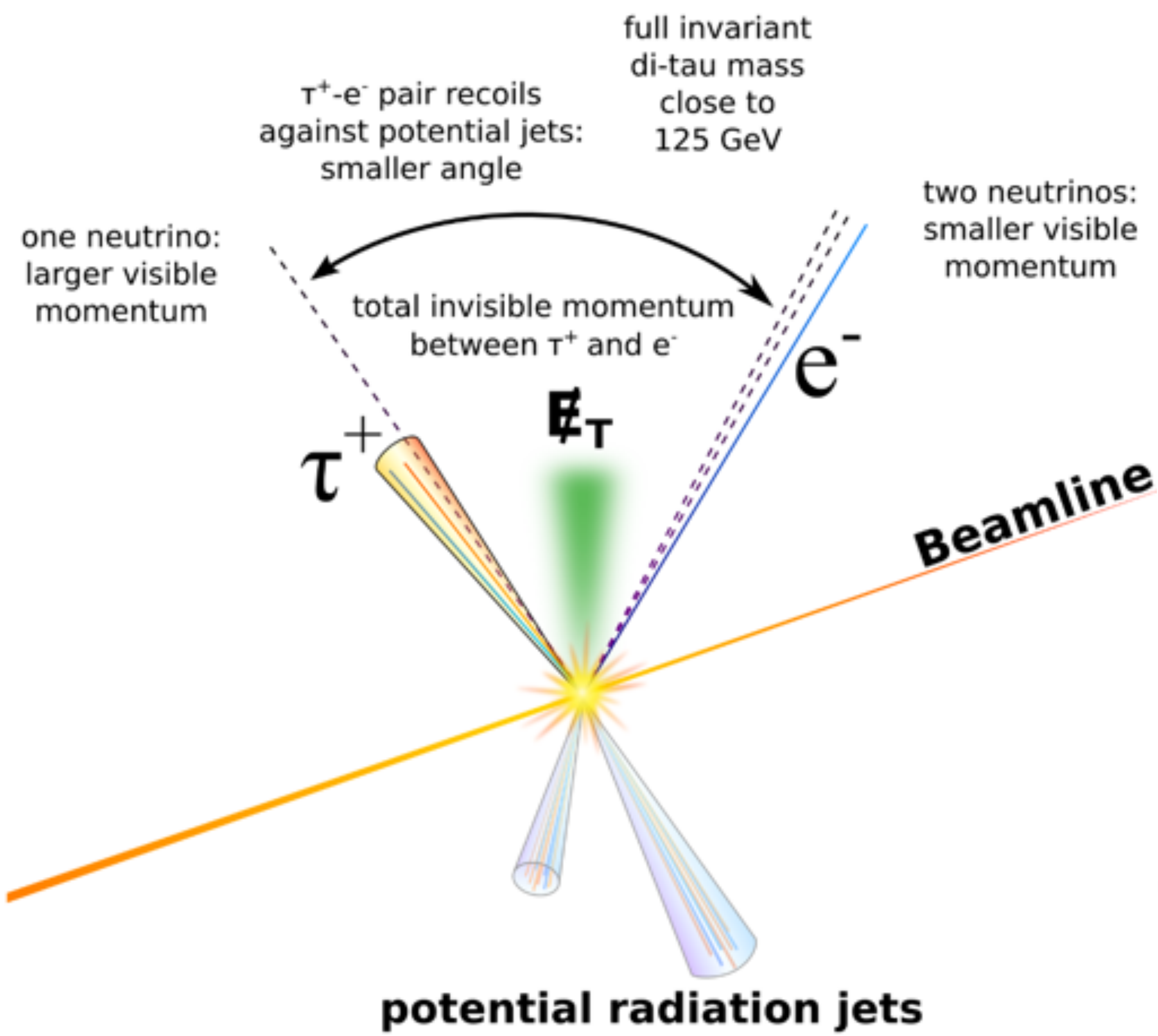
Leading two production modes are used in the main analysis

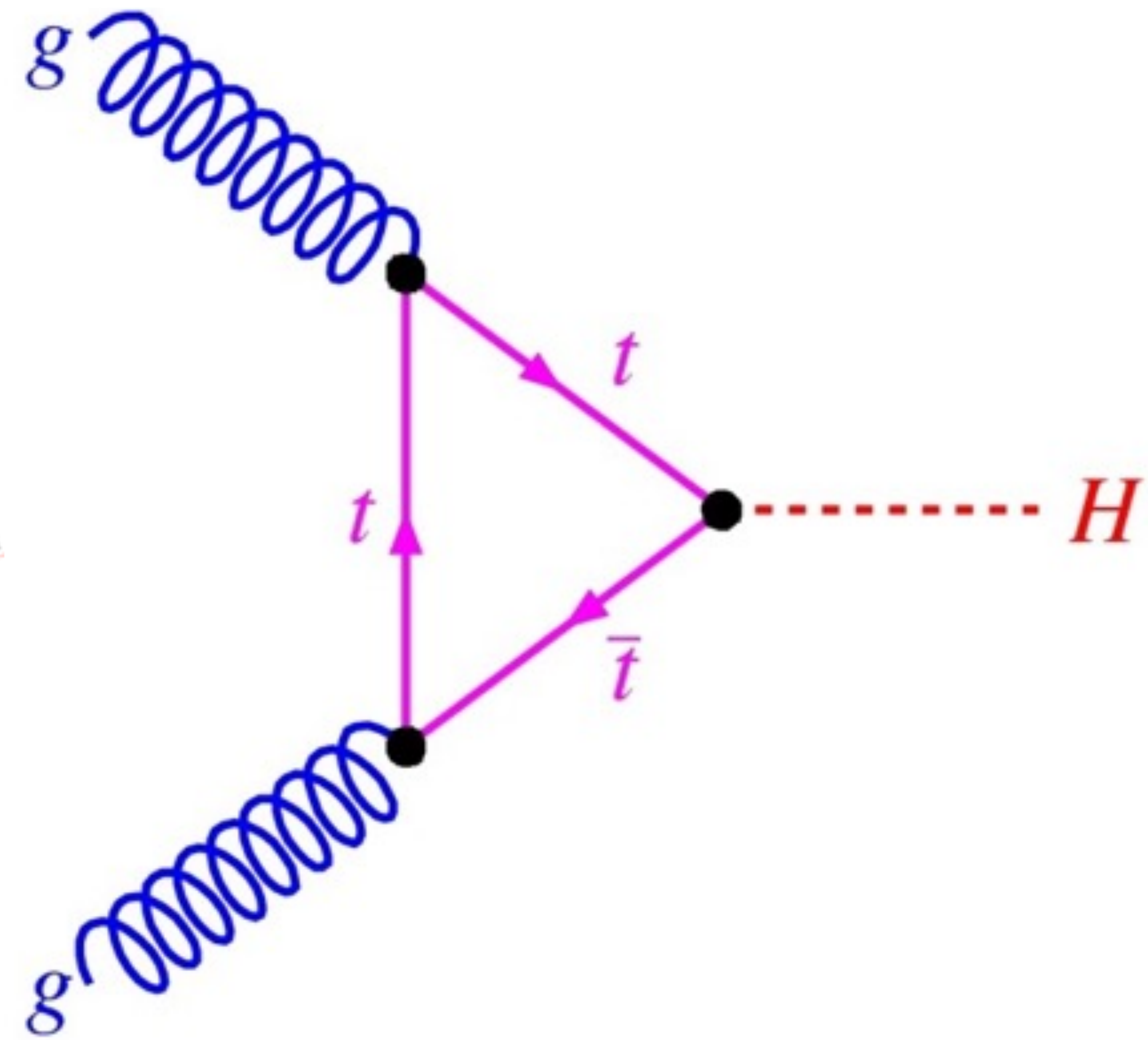
- Gluon-gluon fusion (ggF) and vector boson fusion (VBF)





Signal - ggF





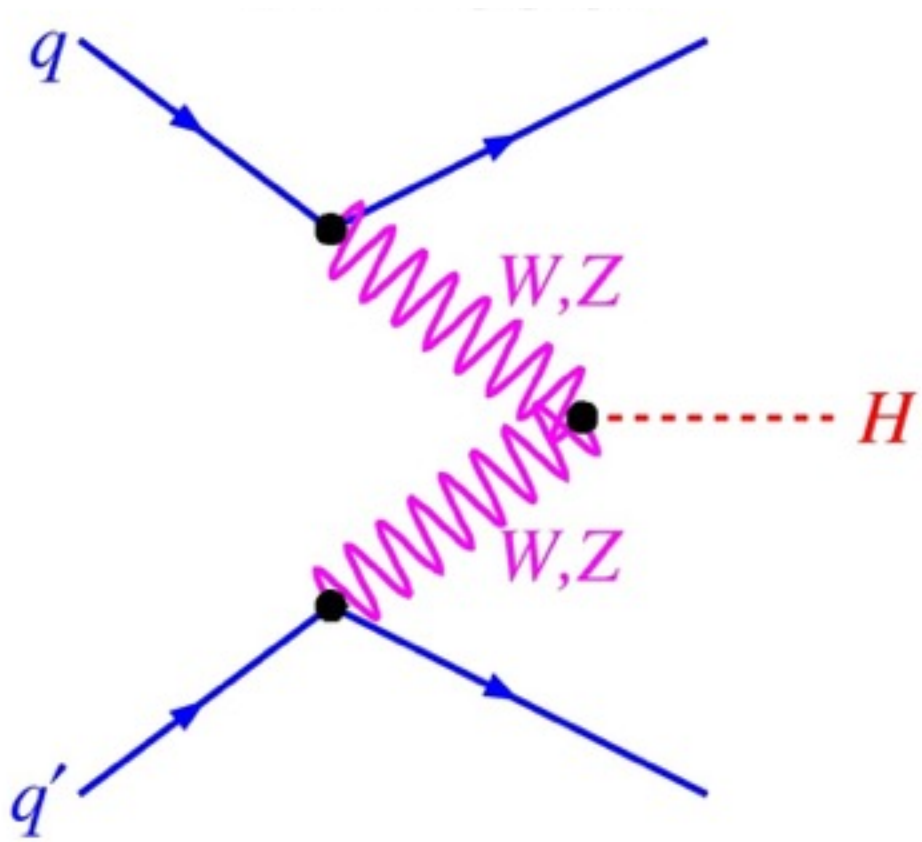
Gluon-gluon fusion contains the **largest cross-section**.

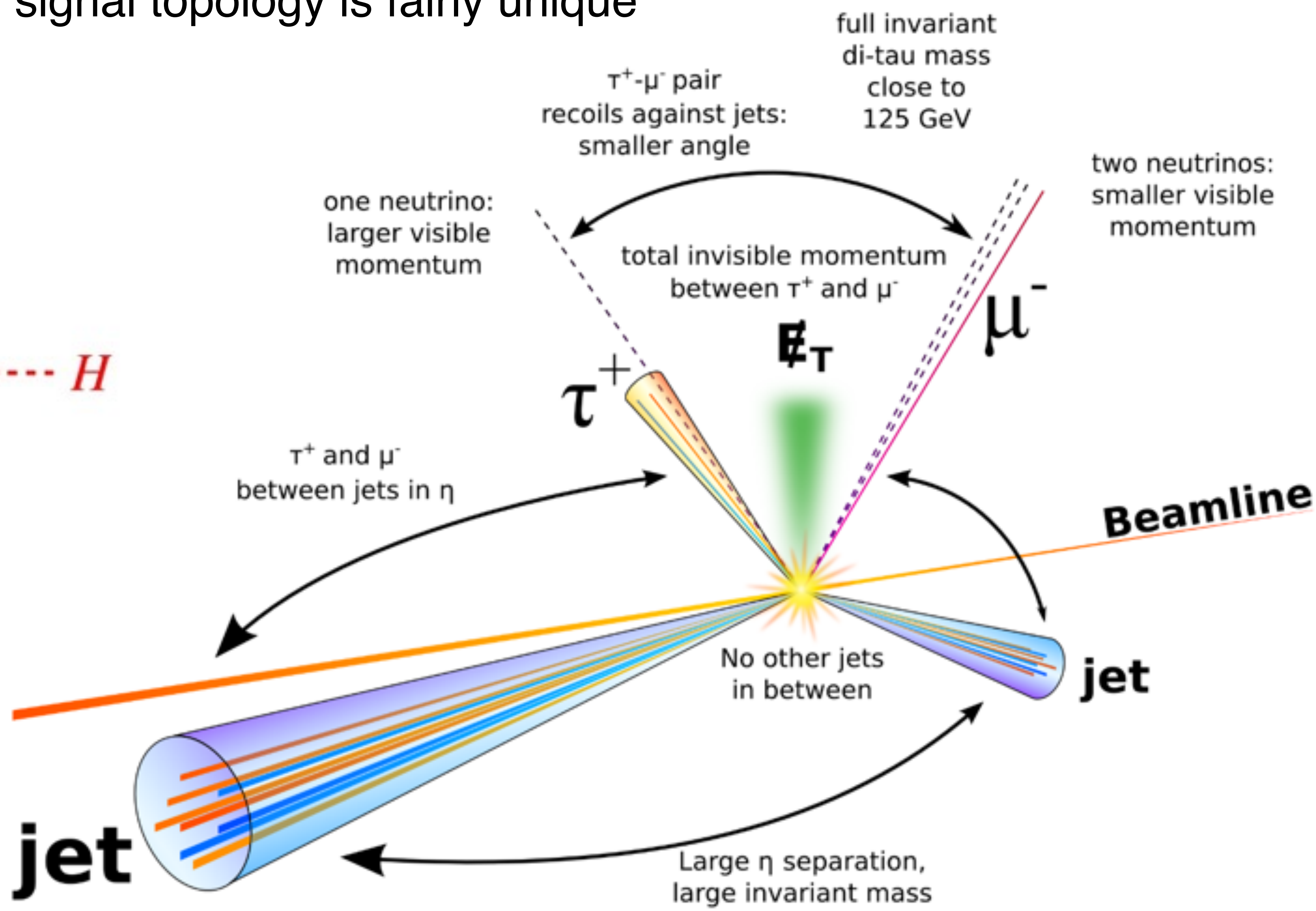
Topology characterised with **large boost in transverse plane recoiling off a jet**

Signal - VBF

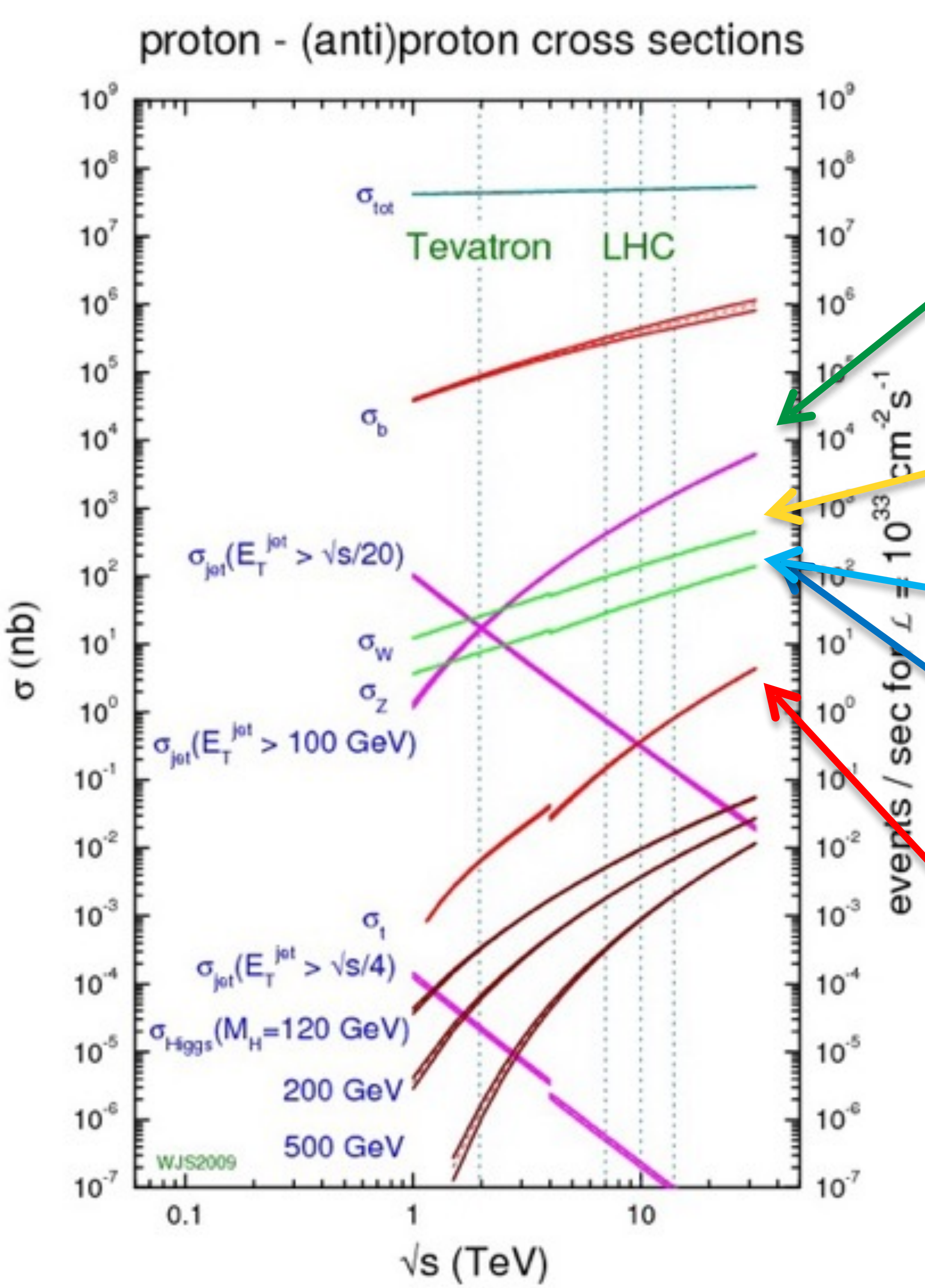
Vector Boson Fusion characterised by **two jets collimated along the beam**

Very sensitive channel as signal topology is fairly unique





Backgrounds



Multijet production increases with higher energy. Jets can be misreconstructed as taus

W+jet production also very large cross-section

Z⟶ll important background for LL and LH (through lepton mis-id)

Z⟶ττ is largest irreducible background

Top backgrounds also important as there is a large multiplicity final state. Important for LL and LH channels only.

Backgrounds

Irreducible:

- Events with **identical prompt final state (Ztt)**

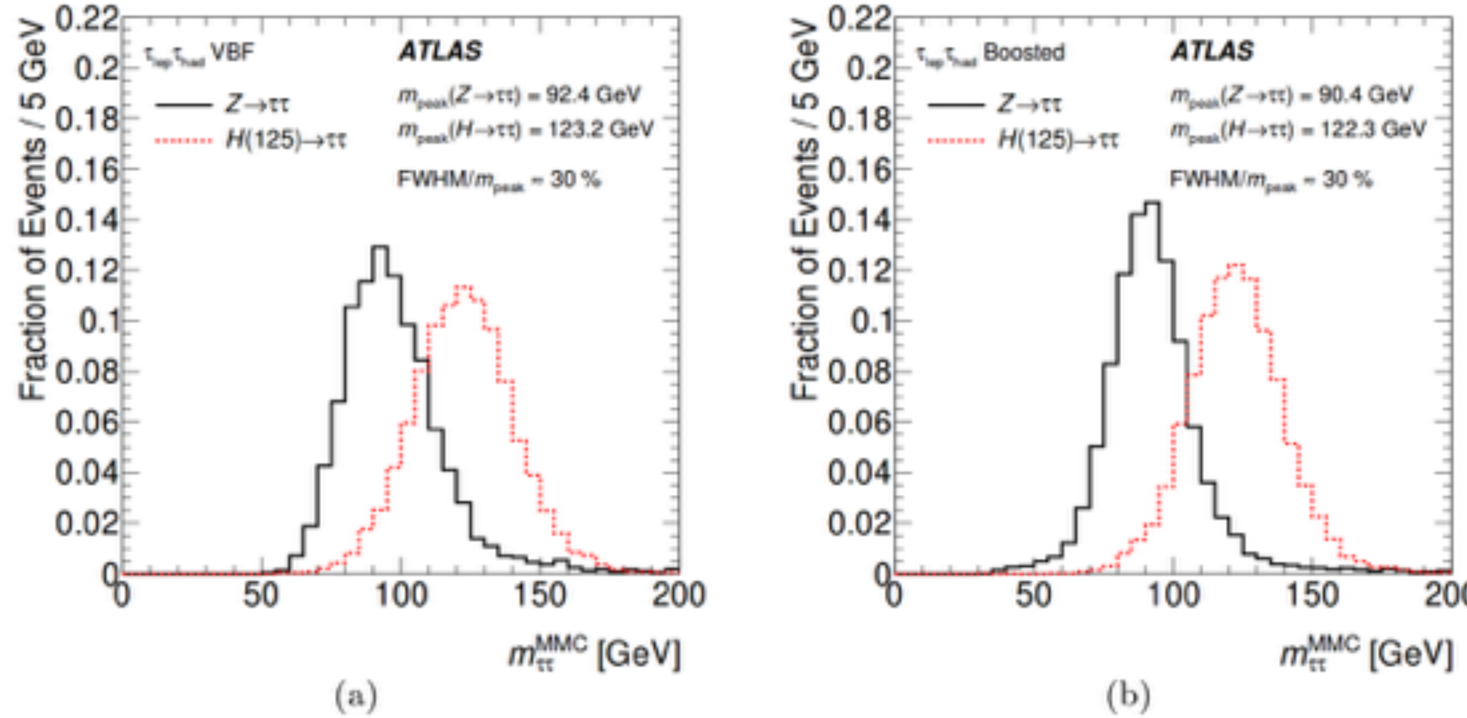
Best discriminator:

- **Mass of boson** decaying to tau pair
- Kinematics of the decay products

Modelling:

- Had a data-driven method in Run I

arxiv:1501.04943



Reducible:

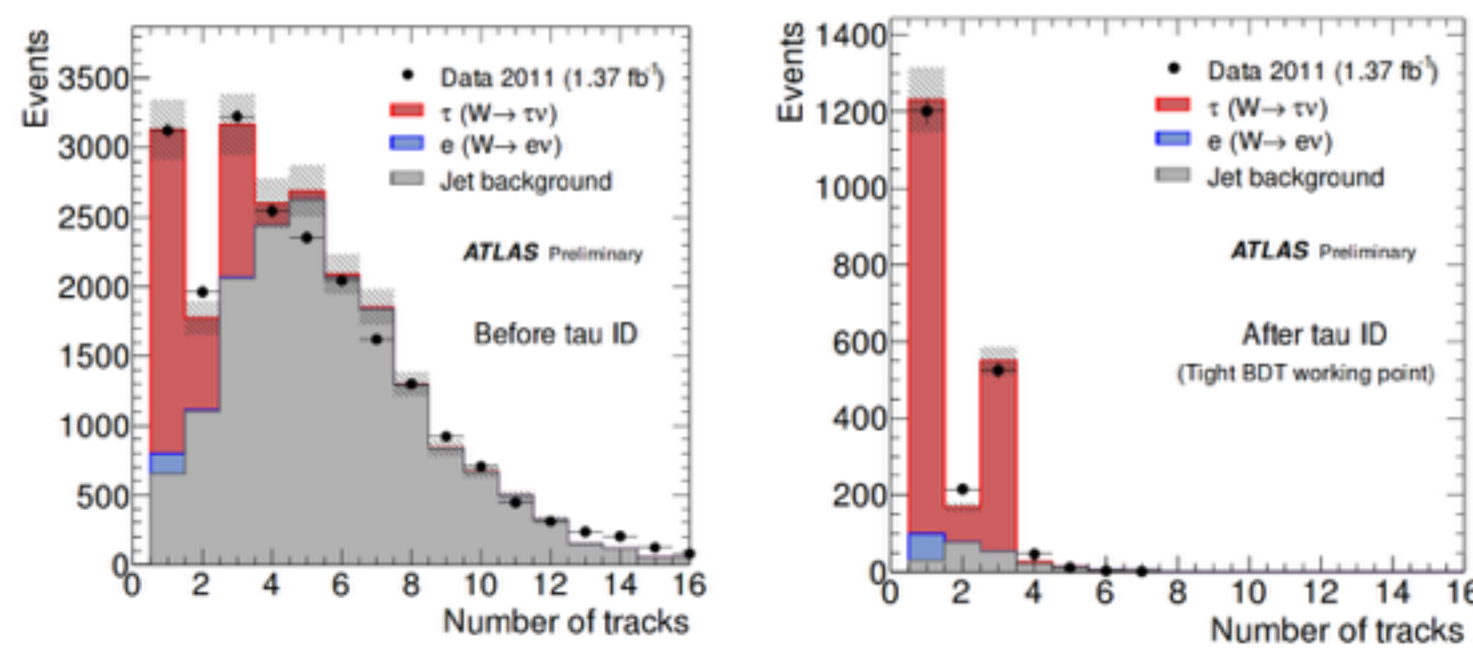
- Events with non-prompt final state (**processes where jet passes tau ID**)

Best discriminator:

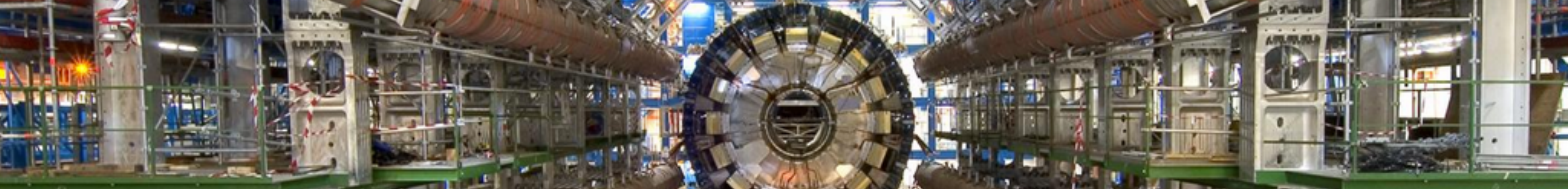
- **Identification requirements** and topology

Modelling:

- Mostly data-driven approaches



arXiv:1412.7086



Part 3:

CP Analysis

HCP Measurement

Search strategy for SM coupling analysis (background estimation, event selection) can be recycled for the CP measurement.

Current status of Higgs CP studies from H⟶VV studies:

- Spin-1 is excluded (from H⟶γγ measurement)
- Pure spin 2 and CP odd scalar disfavoured
- No sensitivity to tree level couplings

H⟶ττ process couple directly through Yukawa interaction:

- Allows for measurement of possible mixed CP states

$$\mathcal{L}_{\mathrm{Y}}^{tau} = -g_\tau \left(\cos\phi_\tau \cdot \bar{\tau}\tau + \sin\phi_\tau \cdot \bar{\tau} i\gamma_5 \tau\right) h$$

$$\Gamma(h_{mix} \to \tau^+\tau^-) \sim 1 - s_{||}^{\tau^+} s_{||}^{\tau^-} + s_{\perp}^{\tau^+} R(2\phi)\, s_{\perp}^{\tau^-}$$

CP sensitive observable

CP of Higgs boson is **encoded in the tau-tau polarisation**

- **Angle between decay planes** is **best observable** to measure this

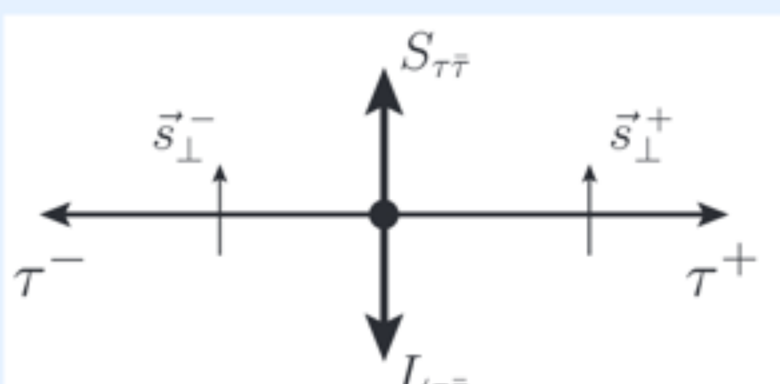
$$h^0$$

CP-even (SM), $\phi_\tau = 0$

$$\mathcal{L}_{h^0\tau\tau} = -g_\tau \cdot \bar{\tau}\tau h$$

$$J^{PC} = 0^{++}$$

$$L_{\tau\bar{\tau}} = 1,\ S_{\tau\bar{\tau}} = 1$$



$$A^0$$

CP-odd, $\phi_\tau = \frac{\pi}{2}$

$$\mathcal{L}_{A^0} = -g_\tau \cdot \bar{\tau} i \gamma_5 \tau h$$

$$J^{PC} = 0^{-+}$$

$$L_{\tau\bar{\tau}} = 0,\ S_{\tau\bar{\tau}} = 0$$

https://indico.desy.de/getFile.py/access?contribId=8&resId=0&materialId=slides&confId=10937

CP sensitive observable

CP of Higgs boson is **encoded in the tau-tau polarisation**

- **Angle between decay planes** is **best observable** to measure this

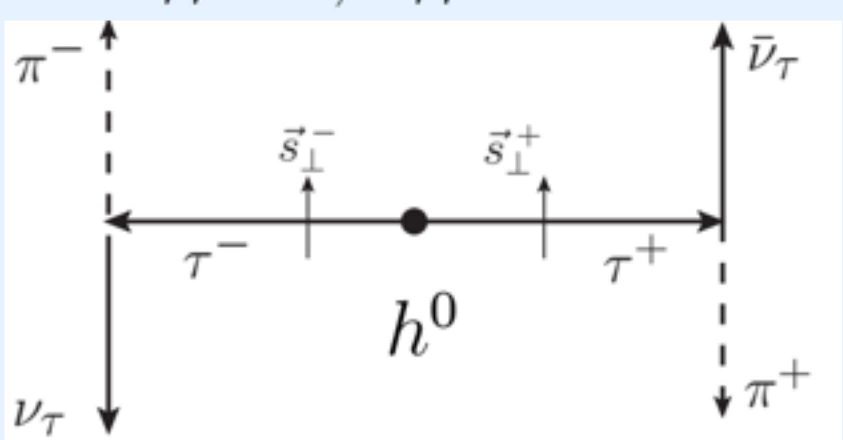
h^0

CP-even (SM), $\phi_\tau = 0$

$$\mathcal{L}_{h^0\tau\tau} = -g_\tau \cdot \bar{\tau}\tau h$$

$$J^{PC} = 0^{++}$$

$$L_{\tau\bar{\tau}} = 1,\ S_{\tau\bar{\tau}} = 1$$



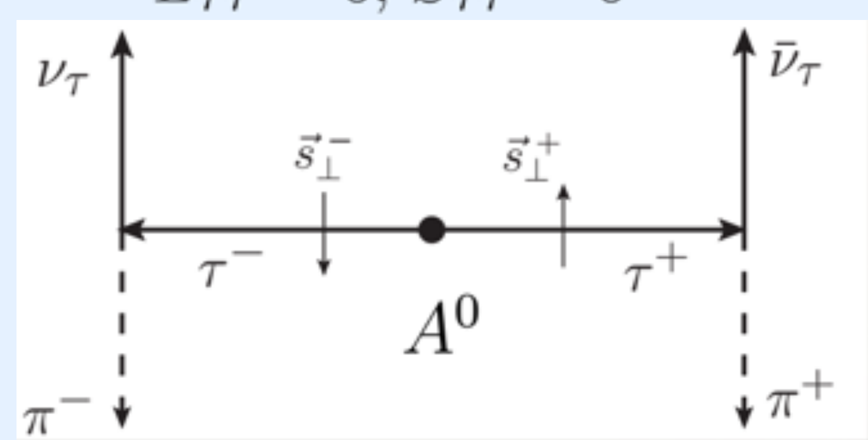
A^0

CP-odd, $\phi_\tau = \frac{\pi}{2}$

$$\mathcal{L}_{A^0} = -g_\tau \cdot \bar{\tau} i\gamma_5 \tau h$$

$$J^{PC} = 0^{-+}$$

$$L_{\tau\bar{\tau}} = 0,\ S_{\tau\bar{\tau}} = 0$$



https://indico.desy.de/getFile.py/access?contribId=8&resId=0&materialId=slides&confId=10937

Impact Parameter Method

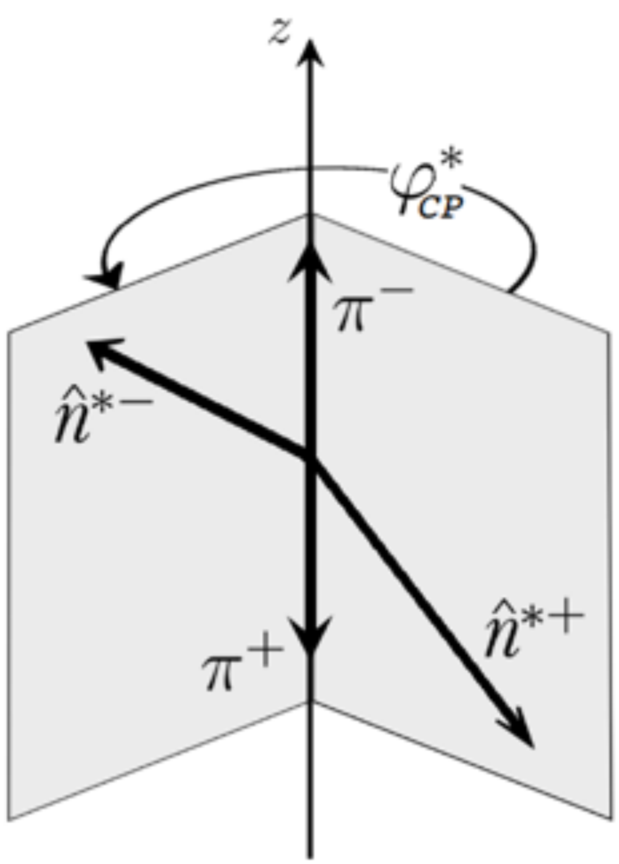
Method :

Approximate the **decay plane** with **impact parameter and the momenta vector**

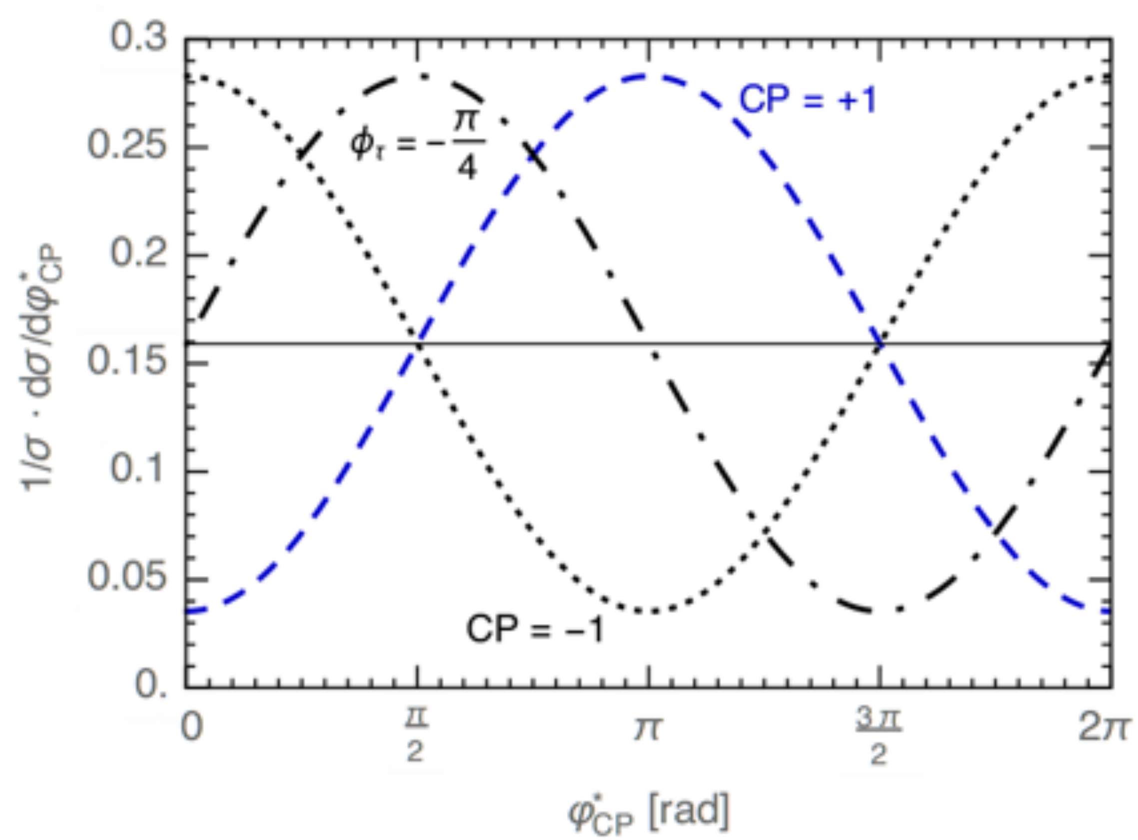
Pro : Can be used across **all one prong decay mode**

Con : Highly dependent on the **resolution of the impact parameter**

$$\varphi^*_{CP} = \begin{cases} \varphi^* & \text{if } \mathcal{O}^*_{CP} \geq 0 , \\ 2\pi - \varphi^* & \text{if } \mathcal{O}^*_{CP} < 0 , \end{cases}$$



arxiv:1510.03850



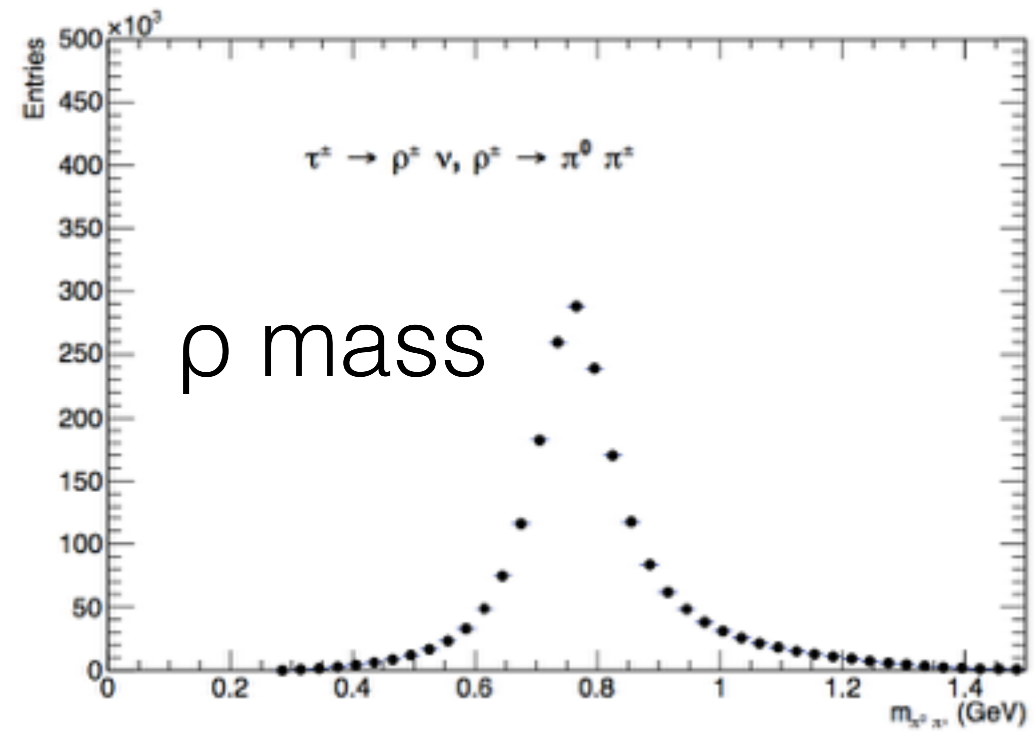
$$\varphi^* = \arccos(\hat{\mathbf{n}}^{*+}_{\perp} \cdot \hat{\mathbf{n}}^{*-}_{\perp}) ,$$

$$\mathcal{O}^*_{CP} = \hat{\mathbf{q}}^*_{-} \cdot (\hat{\mathbf{n}}^{*+}_{\perp} \times \hat{\mathbf{n}}^{*-}_{\perp}) ,$$

Mass distribution

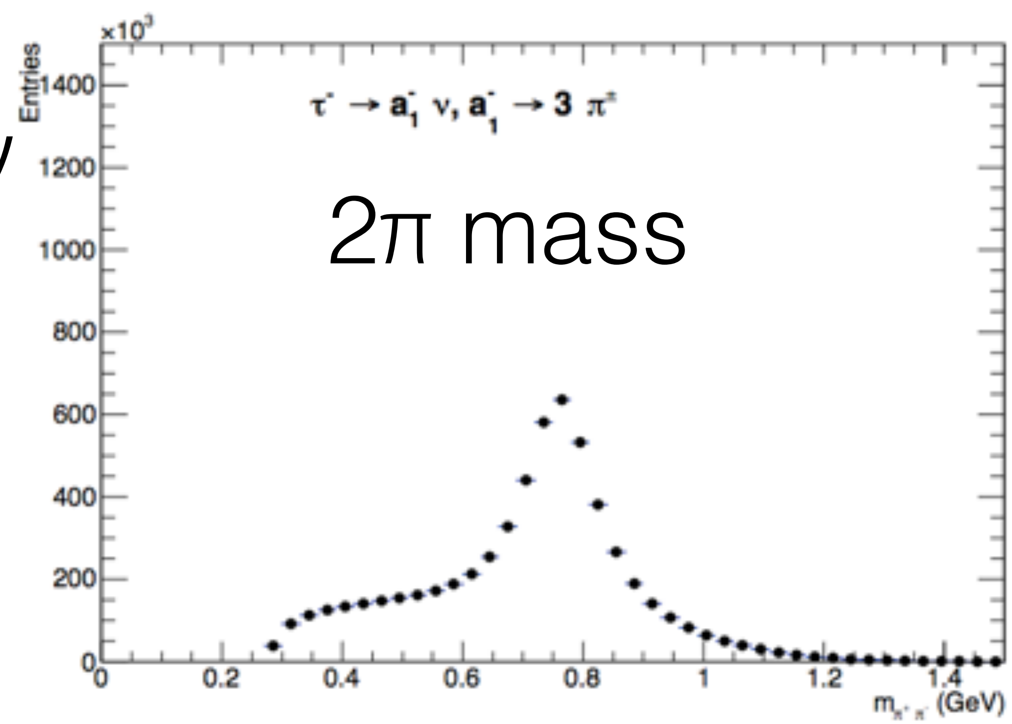
$\tau \longrightarrow \rho \nu$ decay

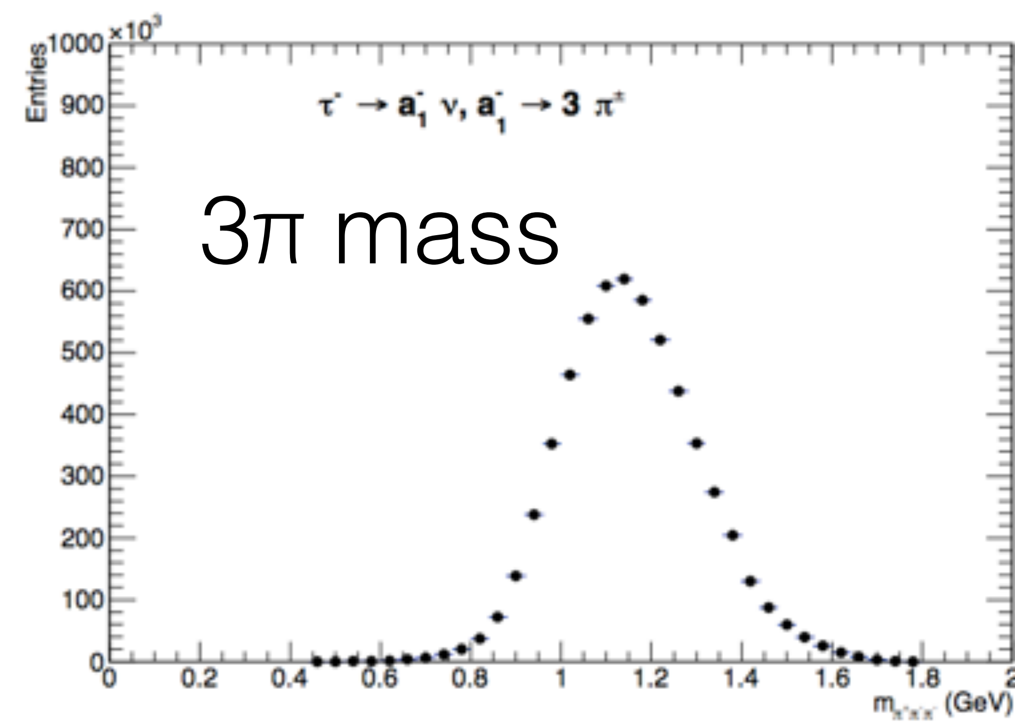
Unique possibility for ρ



$\tau \longrightarrow a_1(\longrightarrow \rho \pi)\nu$ decay

Two interfering cascades ⟶ th. unc. for spin sensitivity





Our NN

Inputs are chosen from the following set:

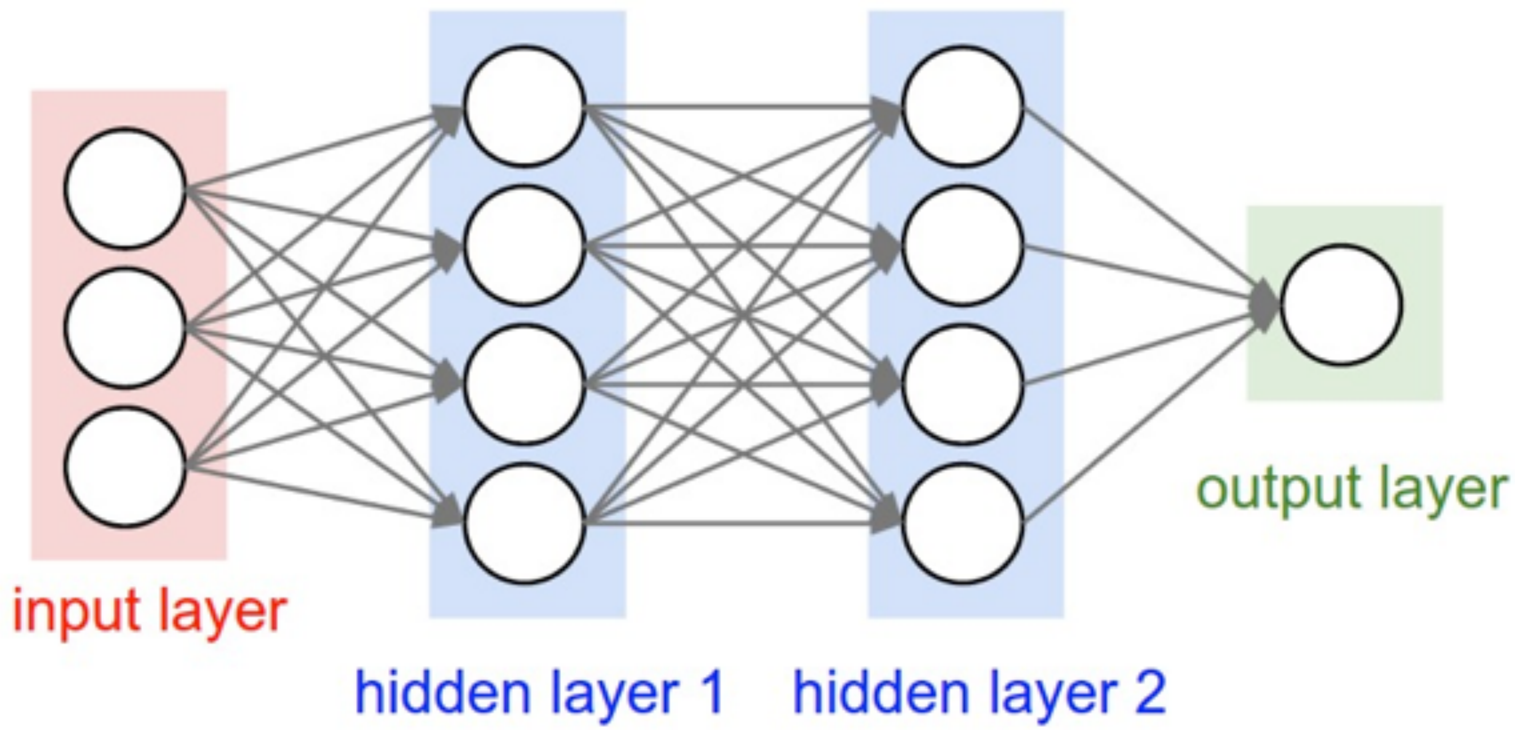
- Acoplanarity angles between particle i and j (φ^*_{ij})
- 4-vectors of visible decay products
- Fractional energy difference between charged and neutral components (y_i, y_j)
- Reconstructed intermediate resonance masses
- …

Combinations of these sets were tested in order to determine an optimal set of inputs. Inputs were also boosted into the reconstructed visible decay product frame.

Our model was built with 6 hidden layers of 300 nodes. Final layer extracted through a sigmoid function.

Three separate networks were trained, one for each combination of intermediate decay mode:

- $\rho\rho$, ρa_1 and $a_1 a_1$



3.2 Metric

We need a quantitative metric to compare different models and approaches. The metric used to compare the models is a weighted Area Under Curve (AUC) [25]. $X = (x_1, ..., x_n)$ is a dataset of interest consisting of n events. $p = (p_1, .., p_n)$ is a vector probabilities returned by neural network model, $p_i = p(x_i \in A) = 1 - p(x_i \in B)$.

The final metric will be computed as follows:

$$\begin{aligned} SCORE(X,p) \quad = \quad & weightedAUC(\\ & [(1, p_1, w_{a_1}), (0, p_1, w_{b_1}), \\ & (1, p_2, w_{a_2}), (0, p_2, w_{b_2}), \\ & ... \\ & (1, p_n, w_{a_n}), (0, p_n, w_{b_n})]) \end{aligned} \tag{6}$$

That is, each event contributes twice to the computation of the score:

[1]Defined in the rest-frame of the appropriate resonance (a_1, ρ, ρ^0) pairs.

9

- $(1, p_i, w_{a_i})$ - corresponds to the case in which model correctly predicts x_i in A with assigned probability p_i. It contributes to the final loss with weight w_{a_i}.
- $(0, p_i, w_{b_i})$ - corresponds to the case in which model incorrectly predicts x_i in A with assigned probability p_i. It contributes to the final loss with weight w_{b_i}.

The first value in a tuple represents the true target (1 means x_i in A), the second is used for ranking events for the purposes of AUC and the last value represents the weight associated with the event.

$SCORE(X, p)$ will return a value of 0.5 for a model that assigns random predictions. Score of 1.0 is reached for perfect separation of the distributions. In practice, perfect score is not achievable on these problems as the distributions are overlapping. It can be shown that the best achievable score is reached when using optimal predictions $(p_i = w_{a_i}/(w_{a_i} + w_{b_i}))$, which corresponds to about 0.782 result (slightly varies by problem). We will call them oracle predictions.

3.3 Model

Deep Neural Networks [24] have been shown to work very well across many different domains, including image classification, machine translation or speech recognition. We will apply similar techniques to our problem. Neural network can be seen as a non-linear map between inputs and outputs. They are often build using chains of matrix multiplications separated by element-wise transformations. We want to distinguish between the two different CP states of Higgs particle so we frame the problems as binary classifications.

The basic architecture used for the problems will contain D-dimensional input (problem dependent) followed by matrix multiplications transforming the input into N-dimensional space with a ReLU non-linearity ($ReLU(x) = max(0,x)$). We apply multiply such transformations to add more expressive power to the model. The largest network used in our experiments had 7 matrix multiplications transforming data points into following sizes

$$D \rightarrow 300 \rightarrow 300 \rightarrow 300 \rightarrow 300 \rightarrow 300 \rightarrow 300 \rightarrow 1 \tag{7}$$

The output is a scalar value that represents an indicator on whether an event looks closer to type `A` or `B`. We would like to represent the output of a neural network as a probability between the two choices. A common way to accomplish this, and also used here, is to use sigmoid non-linearity ($sigmoid(x) = 1/(1+exp(-x))$) on the last layer, which squishes the output into interval $[0.,1]$ and can be interpreted as probability. The metric minimized by the model is negative log likelihood of the true targets under Bernoulli distributions. That corresponds to a loss function equal to:

$$-logp(y|y_h) = -(y == 0) * log(y_h) - (y == 1) * log(1 - y_h), \tag{8}$$

where y_h represents probability outputted by neural network model.

Initially the weights of the matrices are picked randomly and are optimized using a variant of stochastic gradient descent algorithm called Adam [26]. We also used a recent trick called Batch Normalization [27] and Dropout [28] to improve the training of neural network model. Everything was implemented using `TensorFlow` [29], an open-source framework for numerical computations.

[25] G. Arnaud, S. Arnaud, and L. Dominique, http://mlwiki.org/index.php/ROC_Analysis.

[26] D. Kingma and J. Ba, *arXiv:1412.6980* (2014).

[27] S. Ioffe and C. Szegedy, *arXiv:1502.03167* (2015).

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Journal of Machine Learning Research* **15** (2014), no. 1 1929–1958.

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, *Software available from tensorflow. org* **1** (2015).