

Preliminary results on Higgs CPV analysis in tth coupling at $\sqrt{s} = 550 GeV$ LCF

ILD workshop at CERN

Jenny List¹, Nana Ozaki², Aidan Robson³, Junping Tian⁴, Yuyang Zhang^{1,3}

¹DESY Hamburg, ²Mount Holyoke College, ³University of Glasgow, ⁴Univeristy of Tokyo



September 15, 2026

Introduction: Motivation

Why study CPV in the top-Higgs sector

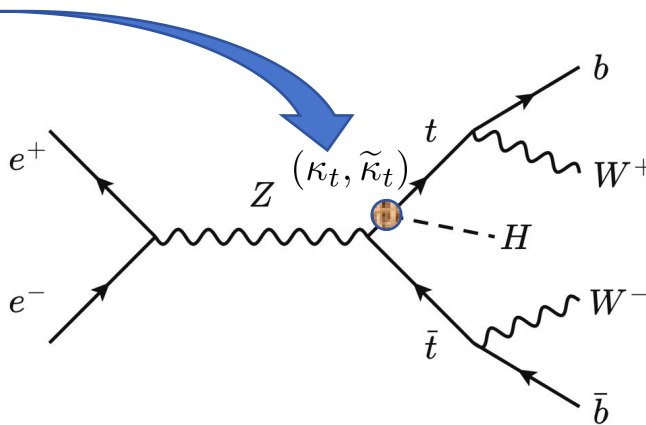
- **Flaw of the SM:** The SM doesn't provide enough CPV to explain the baryon asymmetry of the Universe.
~ 9 orders of magnitude smaller!
- **Physics Chance:** The top quark has the largest Yukawa coupling and a special role in many BSM scenarios.
- **Collider Chance:** Linear e^+e^- collider provides a clean environment with well-defined initial state to probe the CP structure of the top-Higgs interaction through a direct physics process $e^+e^- \rightarrow t\bar{t}h$ at higher energy

Parameterisation of the top-Higgs Yukawa coupling with CPV source

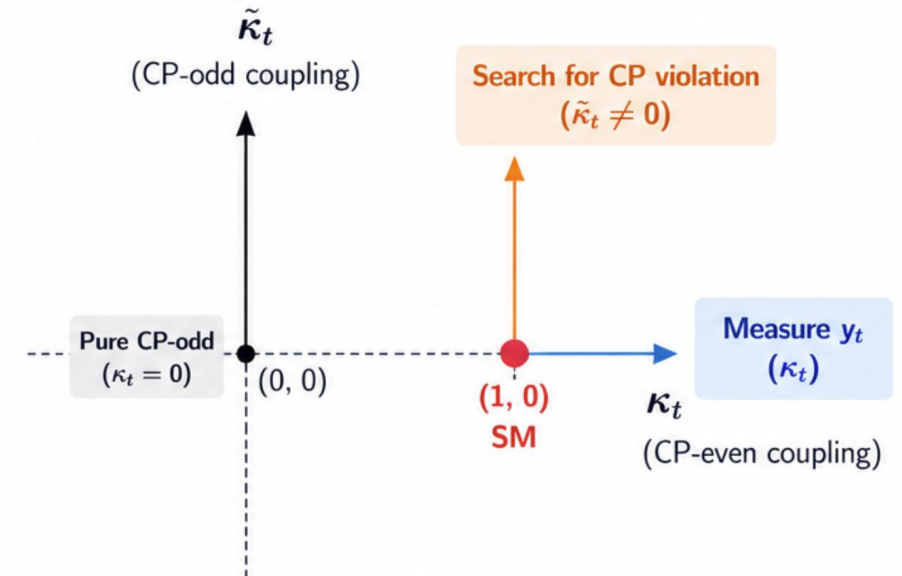
$$\mathcal{L}_{t\bar{t}H} = -\frac{m_t}{v} H \bar{t} \left(\boxed{\kappa_t} + i \boxed{\tilde{\kappa}_t} \gamma_5 \right) t$$

scalar coupling

pseudoscalar coupling



The $(\kappa_t, \tilde{\kappa}_t)$ plane



Introduction: Study Background

The previous precision study at ILC was done by R.Yonamine et al. in 2011

《Measuring the top Yukawa coupling at the ILC at $\sqrt{s} = 500 \text{ GeV}$ 》

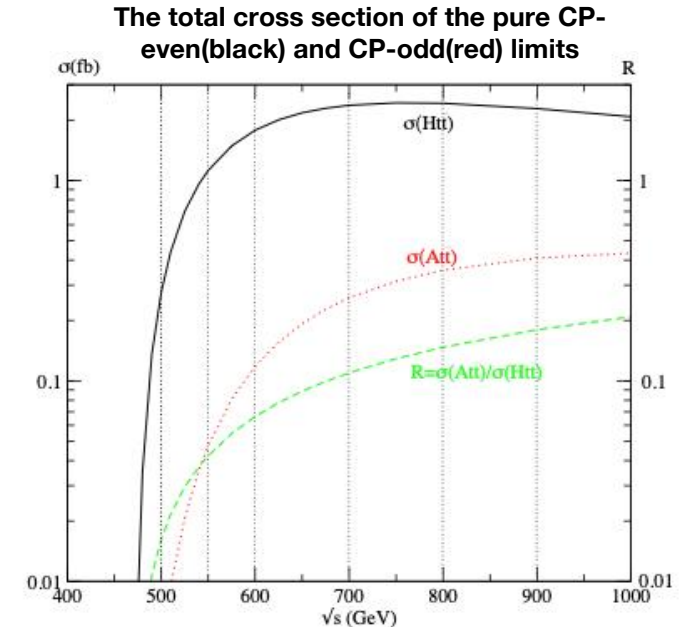
$$\mathcal{L} = 1 \text{ ab}^{-1}, m_H = 120 \text{ GeV}, (P_{e-}, P_{e+}) = (-0.8, +0.3)$$

What happens after 2011?

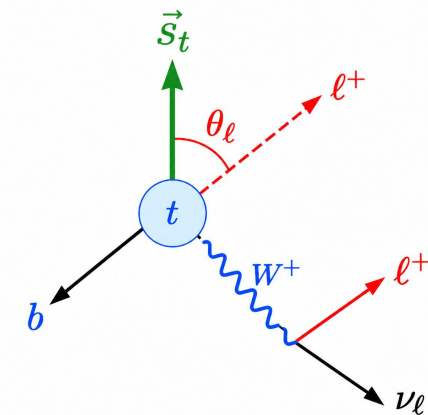
- 125GeV Higgs boson was found in LHC
- New running energy: Study shows $\sqrt{s} = 550 \text{ GeV}$ is a better energy for tth CPV
- New detector design features and updated simulation tools
- New Isolated lepton tagger: Be upgraded to the TMVA based in Marlin
- New jet flavor tagging: ParT implemented ILC allow us to identify 11 class jet flavor
- New kinematic fitting algorithm: SLD neutrino correction, ErrorFlow, 4C fit...
- New S/B separation method: MVA

For CPV observables

- Spin-correlation based observables in this collider environment have been studied at theoretical level
- Machine Learning become widely used to build optimal observable at detector level



K. Hagiwara et al. 1712.09953



Event Generation

$$\mathcal{L}_{t\bar{t}H} = -\frac{m_t}{v} H \bar{t} (\kappa_t + i\tilde{\kappa}_t \gamma_5) t$$

SM Signal: ($\kappa_t = 1, \tilde{\kappa}_t = 0$)

- Physsim: $e^+e^- \rightarrow 6f + h$, 1M for eLpR/eRpL each
- Whizard: $e^+e^- \rightarrow t\bar{t}h$, 50K eLpR
 - *Jet-related reconstruction performance study*

CPV Signal: ($\kappa_t = 1, \tilde{\kappa}_t = 1$)

$$|M|^2 = M_S^2 + M_P^2 + 2\text{Re}(M_S^* M_P)$$

CP-odd term: change sign under CP transformation but contributes nothing to the total cross-section

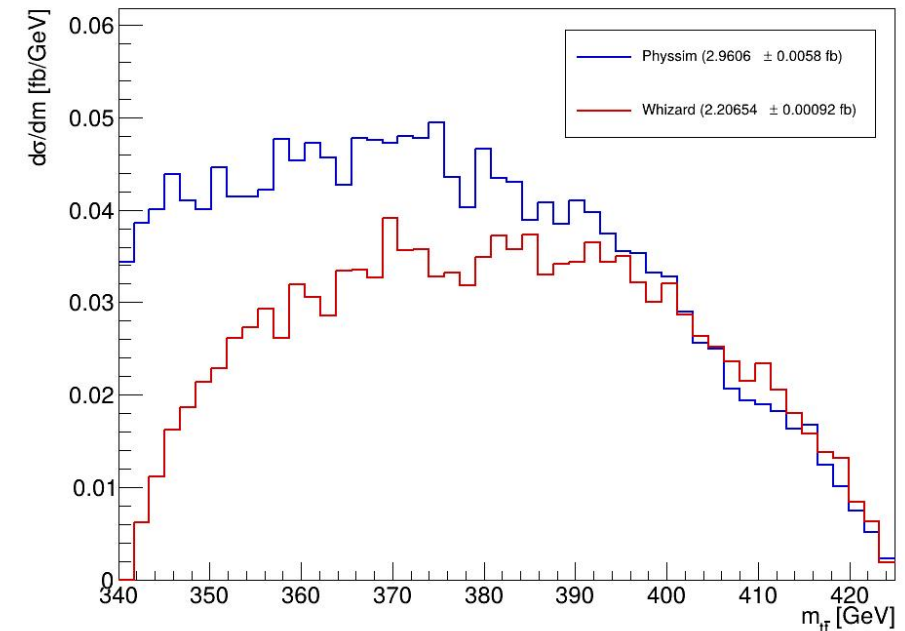
- Physsim: $e^+e^- \rightarrow 6f + h$, 1M for eLpR/eRpL each

Background

- Physsim: $e^+e^- \rightarrow 6f + z$, 1M for eLpR/eRpL each
- Physsim: $e^+e^- \rightarrow 6f + g^*$, $g^* \rightarrow b\bar{b}$, 1M for eLpR/eRpL each
- Whizard: $e^+e^- \rightarrow 6f$, ~8M in total

Detector Simulation: All passed by **SGV**(Fast detector simulation) successfully.

Generator-level $m_{t\bar{t}}$ comparison



Physsim:

- Corrected by the $t\bar{t}$ bound effect
 $\sqrt{s} \sim 2m_t + m_H \sim 475 \text{ GeV}$
- Resonant only ($m_{bW} \sim m_{top}$)
- $e^+e^- \rightarrow Z^* + h$, $Z^* \rightarrow t\bar{t}$ is also included

Semileptonic Selection(electron,muon)

Pre-selected by the total muon and electron number (BDT Isolated lepton Tagger)

	tth total	tth, truth semi-lep(e,mu)	ttz	ttg	6f
IsoMu+IsoE==1	37%	94%	34%	36.4%	31%
Isolep==1 (old Finder)	-	88%	-	-	-

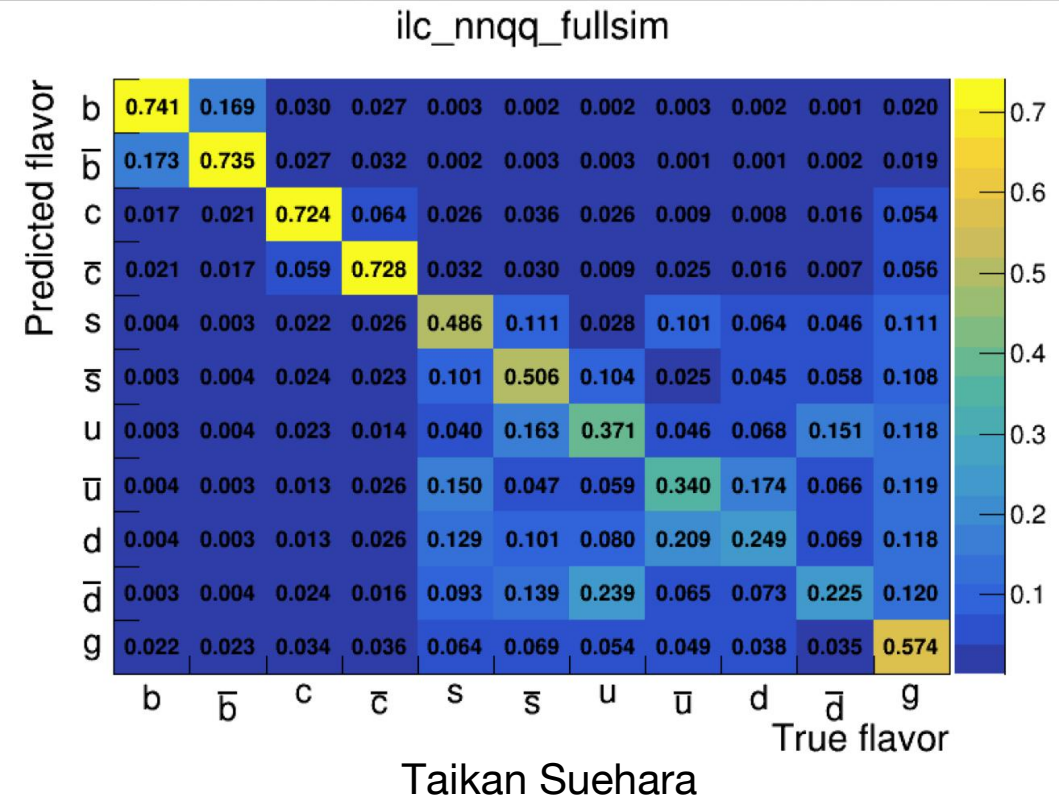
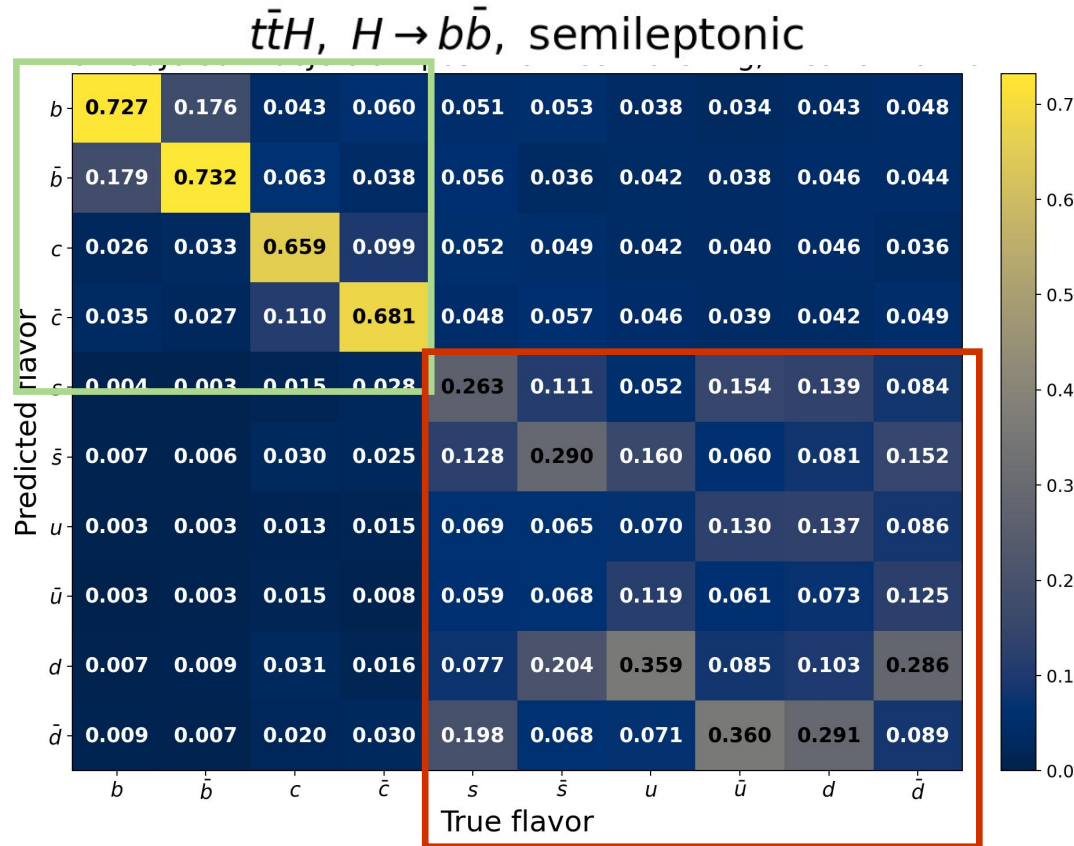
In the selected tth semileptonic events

	tth, truth semi- lep(e,mu)	tth, truth semi-lep(tau)	tth, truth hadronic	tth, truth dileptonic
BDT Tagger	77%	10%	1%	12%
Cut Finder	71%	11%	7%	11%

- The BDT tagger improved the efficiency by 6%
- BDT tagger has a higher ability on excluding the hardonic background, cannot perform better on excluding tau.

Jet Flavor Tagging Performance

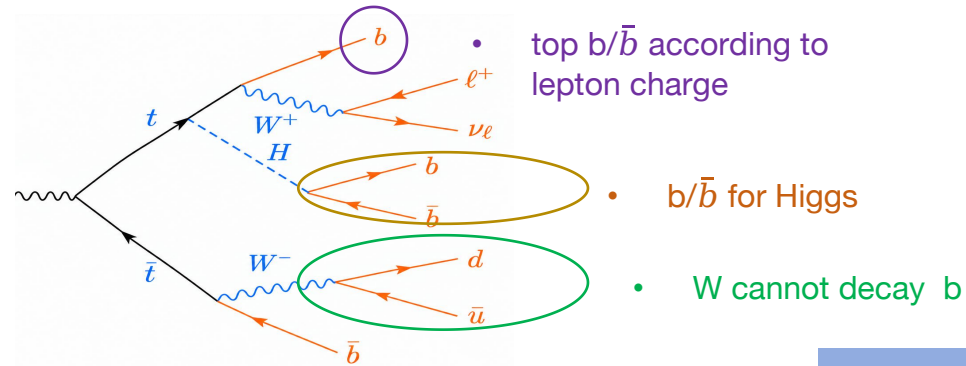
- In truth $t\bar{t}H$, $h \rightarrow b\bar{b}$, semileptonic channel, 6 Jets are clustered by Durham algorithm
- The overall performance of ParT is much better than LCFIplus (b/c-tag distribution...see backup)



- The high accuracy on distinguishing $b/\bar{b}$, $c/\bar{c}$ significantly benefits the jet assignment and signal selection
- The lower accuracy on the light jets cause difficulty on the construction of the CP observables, where need to distinguish up/down-type quark.

Jet Assignment And Kinematic Fit

Event flavor score q_{flavor}



$$q_{\text{flavor}} = S_W + S_H + S_{\text{top}}$$

$$S_W = - [\ln(p_{\text{non-b}}^{w1}) + \ln(p_{\text{non-b}}^{w2})]$$

$$S_H = - [\ln(p_b^{H1} p_{\bar{b}}^{H2} + p_{\bar{b}}^{H1} p_b^{H2})]$$

$$S_{\text{top}} = \begin{cases} -[\ln(p_b^{\text{lep-b}}) + \ln(p_{\bar{b}}^{\text{had-b}})] & q_l > 0 \\ -[\ln(p_{\bar{b}}^{\text{lep-b}}) + \ln(p_b^{\text{had-b}})] & q_l < 0 \end{cases}$$

Kinematic Fit:

- ErrorFlow-based jet uncertainties combined detector resolution, PFA confusion and neutrino correction
- SLD neutrino correction
- 4C: E, px, py, pz conservation + ISR
- Soft W/t/H mass constraints

$$\chi_{\text{measure}}^2 = \sum_i (\mathbf{y}_i^{\text{fit}} - \mathbf{y}_i^{\text{meas}})^T V_i^{-1} (\mathbf{y}_i^{\text{fit}} - \mathbf{y}_i^{\text{meas}})$$

The event reconstruction score q_{reco}

$$\chi_{\text{fit}}^2 = \chi_{\text{measure}}^2 + \chi_{\text{mass}}^2$$

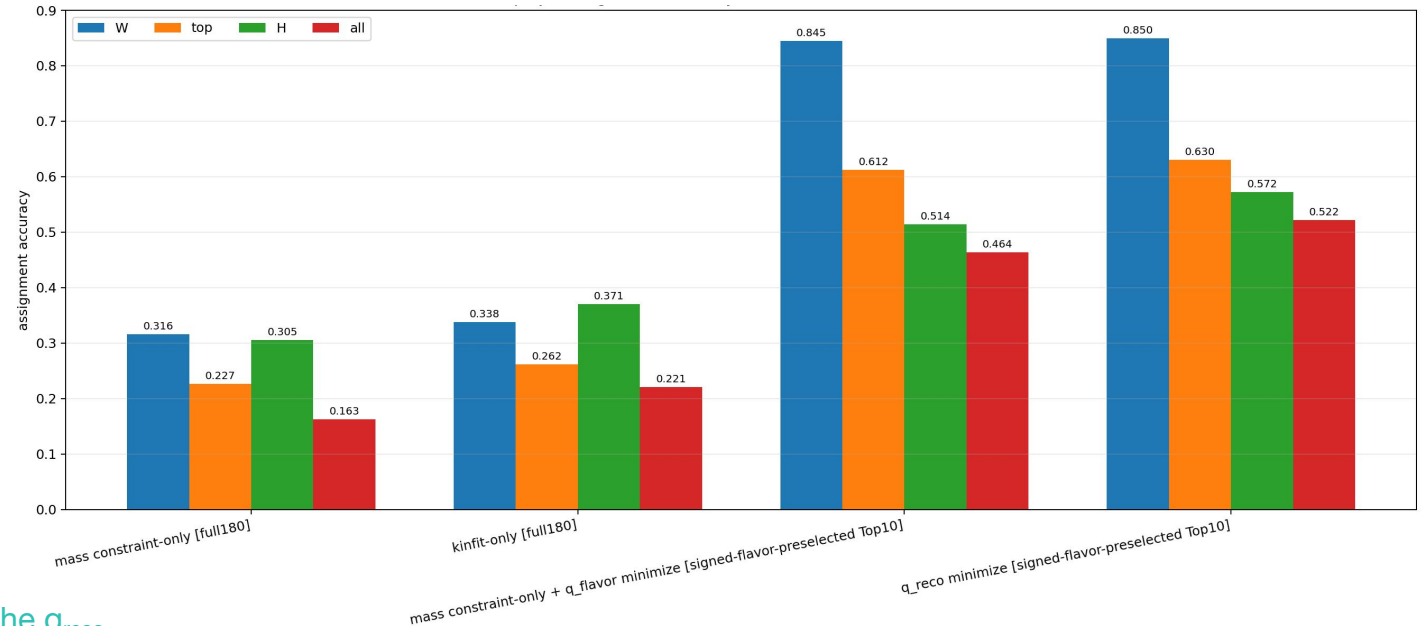
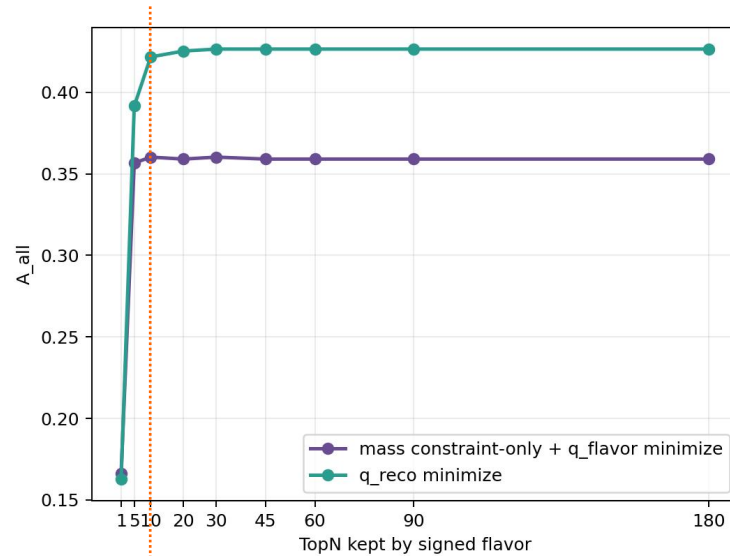
$$\chi_{\text{mass}}^2 = \left(\frac{m_W^{\text{fit}} - M_W}{\sigma_W} \right)^2 + \left(\frac{m_{t,\text{had}}^{\text{fit}} - M_t}{\sigma_{t,\text{had}}} \right)^2 + \left(\frac{m_H^{\text{fit}} - M_H}{\sigma_H} \right)^2 + \left(\frac{m_{t,\text{lep}}^{\text{fit}} - M_t}{\sigma_{t,\text{lep}}} \right)^2$$

$$q_{\text{reco}} = \ln(1 + \chi_{\text{fit}}^2) + 0.3q_{\text{flavor}}$$

Jet Assignment

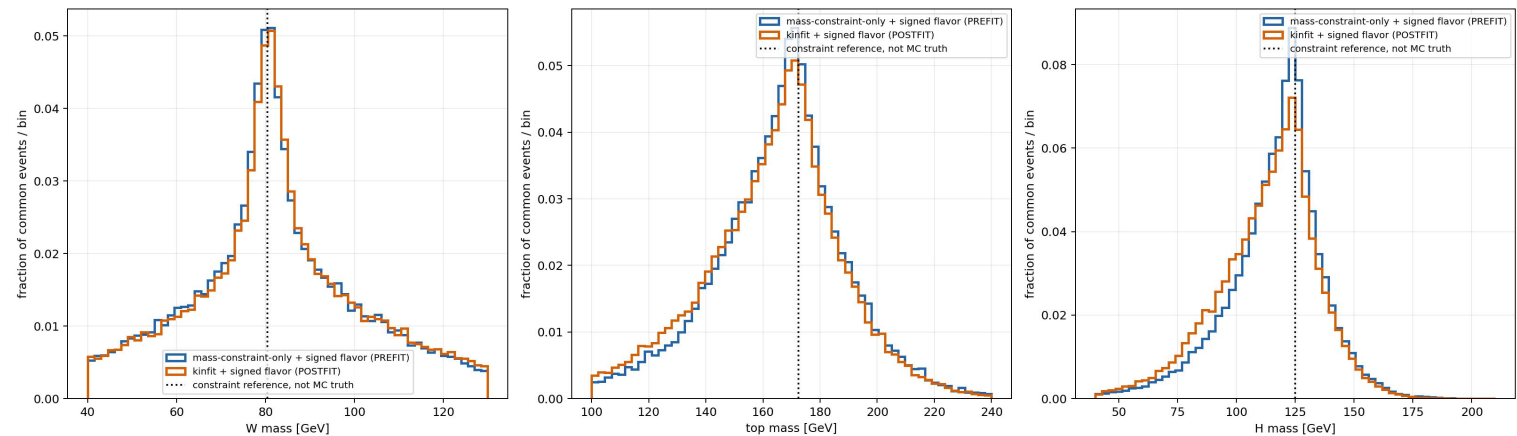
- Reduce the number of the jet permutation
- Improve the accuracy of the assignment

Rank the 180 permutation by the q_{flavor}



- Initially, we choose 10 top q_{flavor} assignments to minimize the q_{reco}

- Mass reconstruction(Physsim)

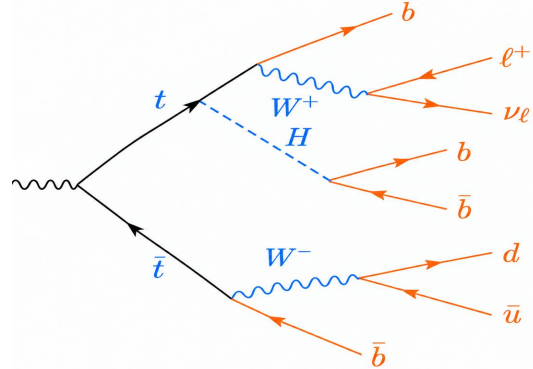


*The TopN scan was tested by 830 selected whizard events, the whole accuracy was plotted by the whole ~4000 whizard event

Cross-Road

Physsim/Whizard → SGV → Semileptonic selection → Jet clustering+flavor tagging → KinFit+Jet Assignment

Reconstructed
Physics Objects
in **semileptonic
channel**



Event
level/kinamatic fit
quantities

$m_W, m_H, m_t,$
 $\chi^2_{fit}, p_{fit}, q_{flavor},$
 $q_{reco}...$

Precise measurement of the coupling strength

- Fix $\tilde{\kappa}_t = 0$, measure κ_t
- Optimise signal-background separation

$$Z_{S/B} = \frac{S_{SM}^{tth}}{\sqrt{S_{SM}^{tth} + B}}$$

- MVA S/B classification → $q_{SB} \rightarrow$ Event Selection

Exploit CP-sensitive information

- Fix $\kappa_t = 1$, constraint $\tilde{\kappa}_t$
- Consider the CPV contribution dependent on the coefficient linearly

$$v_i(\tilde{\kappa}_t) \approx v_{SM,i}^{tth} + \tilde{\kappa}_t v_{cpv,i} + \mathcal{O}(\tilde{\kappa}_t^2)$$

Optimise information

$$I = \sum_i \frac{v_{cpv,i}^2}{v_{SM,i}^{tth}}$$

MVA Signal Background Separation

Model Training

Binary class, XGBoost:

Signal: $t\bar{t}h, h \rightarrow b\bar{b}$ (SM)

Background:

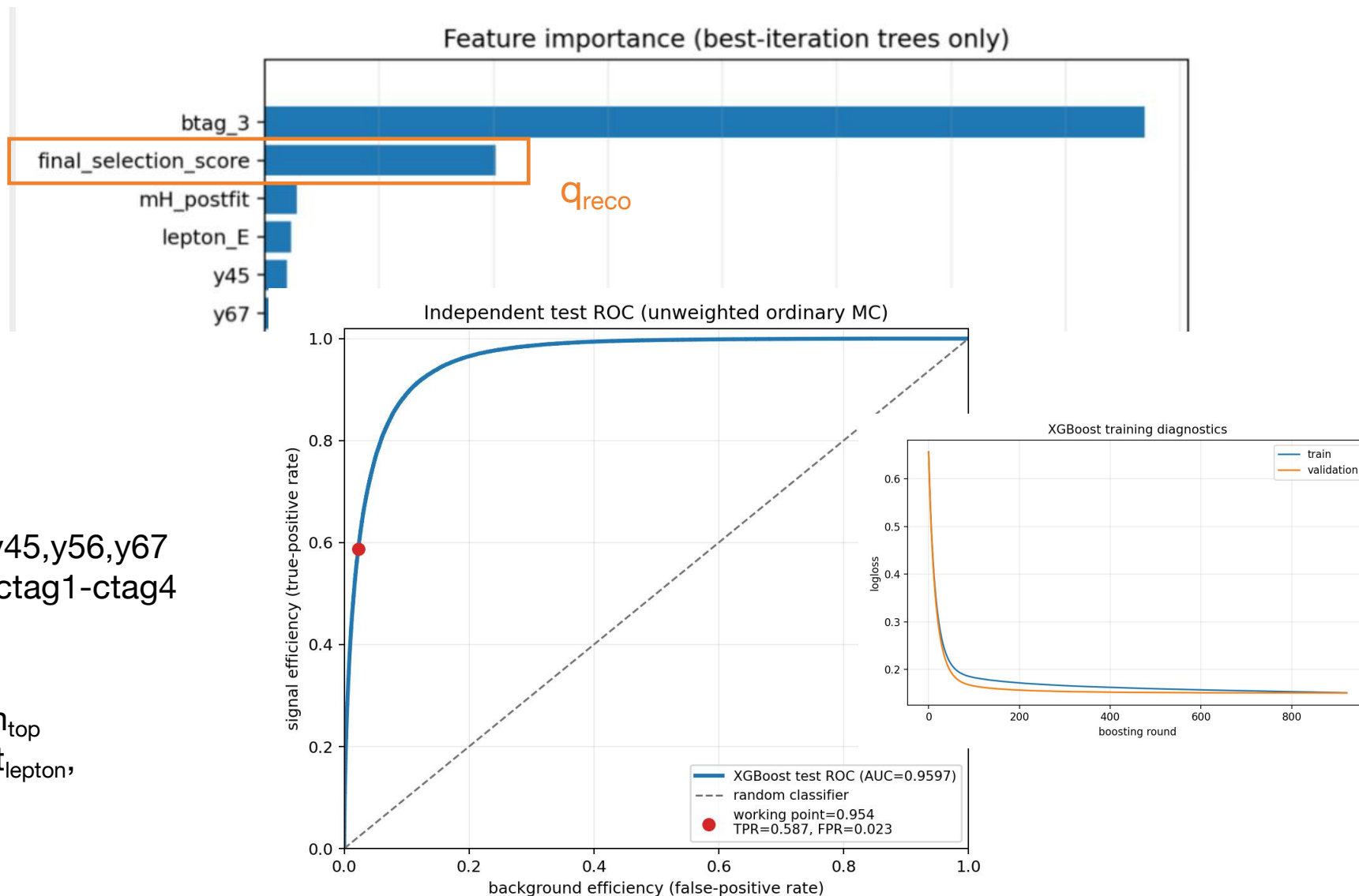
- $6f(t\bar{t})$: 6q+4q2l
- $t\bar{t}h, h \rightarrow non - b\bar{b}$
- $ttz, ttg^*(g^* \rightarrow b\bar{b})$

Baseline Input features

- Jet clustering Durham ymerge: y45,y56,y67
- Jet Flavor Tagging: btag1-btag4, ctag1-ctag4
- Jet assignment and kinfit score:
 $\chi^2_{fit}, p_{fit}, q_{flavor}, q_{reco}$

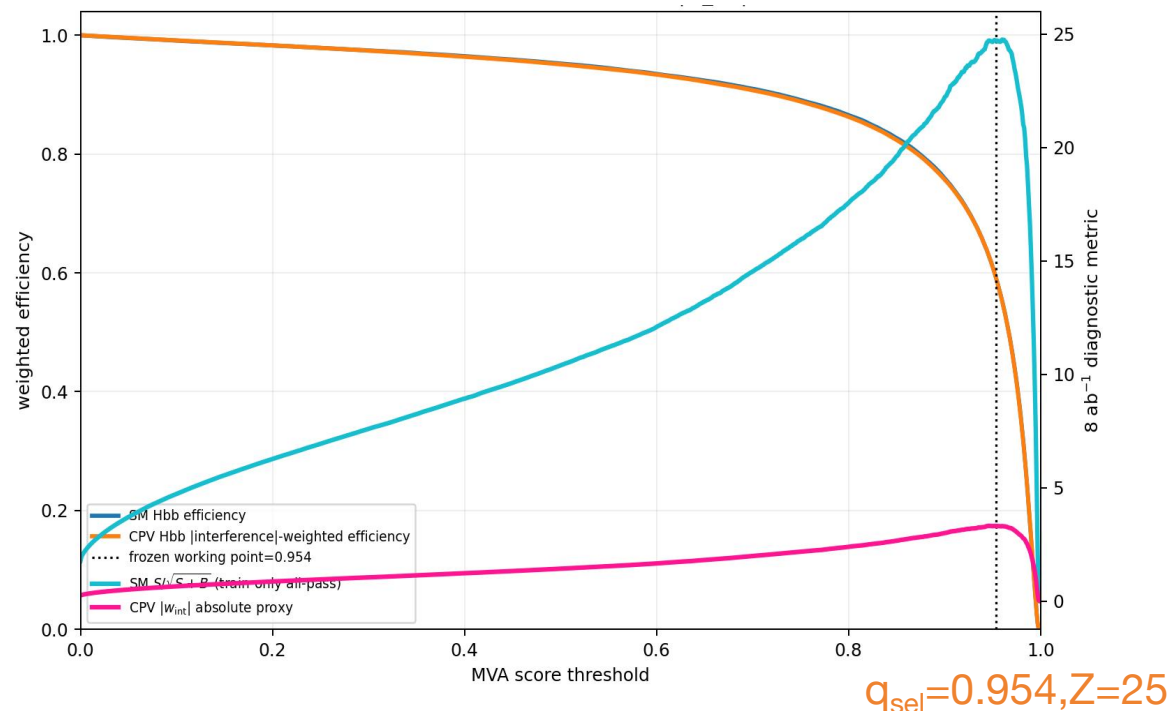
Invariant Mass post kinfit: m_H, m_W, m_{top}

- Isolated lepton: $E_{lepton}, \theta_{lepton}, pt_{lepton},$
lepton_charge



Signal Significance

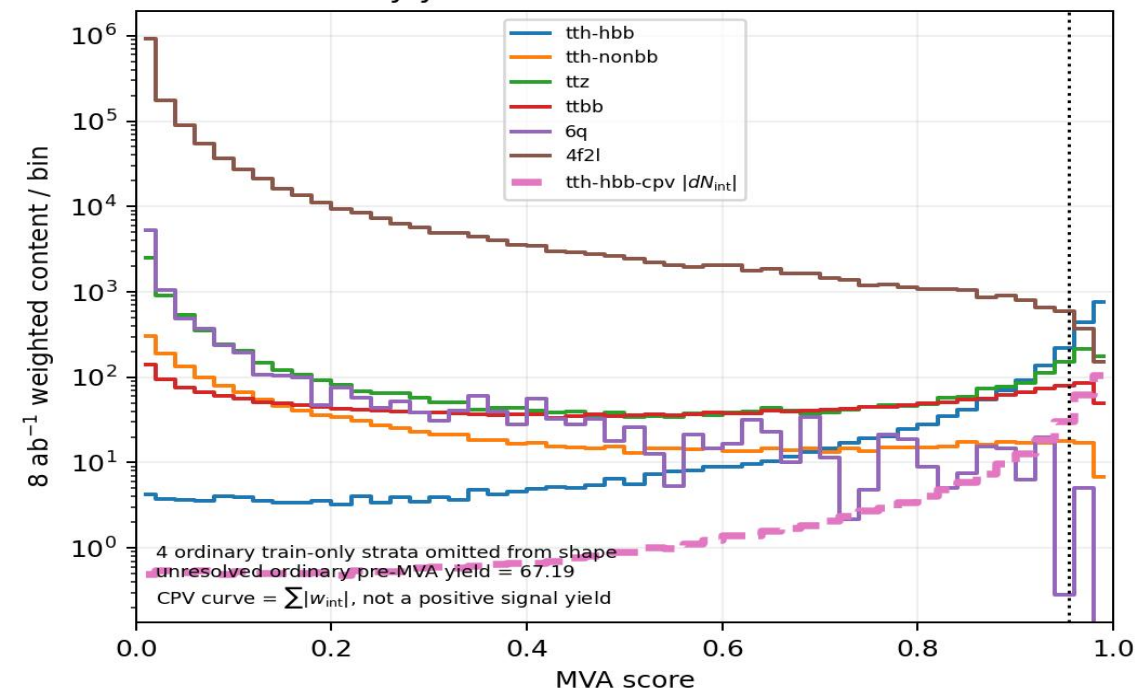
Efficiency vs Significance



Selection Efficiency and physical yield (LCF polarization, 8 ab^{-1} in total)

sample	tth-sm/tth-hbb	tth-cpv	tth-sm/tth-nonbb	ttz	ttbb	6q	4f2l
Efficiency	58.68%	58.87%	1.67%	5.44%	6.29	0.7%	0.4%
Yield	1292	177	30	445	176	5	692

MVA score distribution of the signal and background



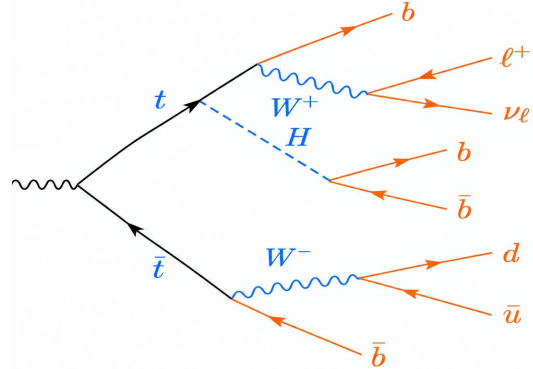
Compare with the 2011 study

- Our results in semileptonic channel
 $Z = 25 \Rightarrow |\Delta \kappa_t / \kappa_t| = 2\%$
- Generalize the 2011 study in "semileptonic channel+ hadronic channel" to 8ab, 550GeV in
2503.19983: $|\Delta \kappa_t / \kappa_t| = 1.9\%$
- Expect higher accuracy in hadronic channel, so the total accuracy will be **smaller than 1.4%**

Cross-Road

Physsim/Whizard → SGV → Semileptonic selection → Jet clustering+flavor tagging → KinFit+Jet Assignment

Reconstructed
Physics Objects
in **semileptonic
channel**



Event
level/kinamatic fit
quantities

$m_W, m_H, m_t,$
 $\chi^2_{fit}, p_{fit}, q_{flavor},$
 $q_{reco}...$

Precise measurement of the coupling strength

- Fix $\tilde{\kappa}_t = 0$, measure κ_t
- Optimise signal-background separation

$$Z_{S/B} = \frac{S_{SM}^{tth}}{\sqrt{S_{SM}^{tth} + B}}$$

- MVA S/B classification → $q_{SB} \rightarrow$ Event Selection
- Got $|\Delta\kappa_t/\kappa_t|=2\%$ at in semi-leptonic channel

Exploit CP-sensitive information(eLpR, 8ab⁻¹)

- Fix $\kappa_t = 1$, constraint $\tilde{\kappa}_t$
- Consider the CPV contribution dependent on the coefficient linearly

$$v_i(\tilde{\kappa}_t) \approx v_{SM,i}^{tth} + \tilde{\kappa}_t v_{cpv,i} + \mathcal{O}(\tilde{\kappa}_t^2)$$

Optimise fisher information

$$I = \sum_i \frac{v_{cpv,i}^2}{v_{SM,i}^{tth}} \quad \tilde{\kappa}_t \sim \frac{1}{\sqrt{I}}$$

CPV observable

CP-odd term in the $|M(x)|^2$:

$$2\text{Re}(M(x)_S^* M(x)_P) \longrightarrow$$

CP $x \rightarrow \bar{x}$:

- particle $\rightarrow$ anti-particle
- $\vec{p} \rightarrow -\vec{p}$

$$\longrightarrow -2\text{Re}(M(\bar{x})_S^* M(\bar{x})_P)$$

CPV observable

CP-odd term in the $|M(x)|^2$:

$$2\text{Re}(M(x)_S^* M(x)_P)$$



$$\mathcal{O}_{cp}(x)$$



CP $x \rightarrow \bar{x}$:

- particle $\rightarrow$ anti-particle
- $\vec{p} \rightarrow -\vec{p}$

$$-2\text{Re}(M(\bar{x})_S^* M(\bar{x})_P)$$



$$-\mathcal{O}_{cp}(\bar{x})$$



CPV observable

$$2\text{Re}(M(x)_S^* M(x)_P)$$



$$\mathcal{O}_{cp}(x)$$



CP $x \rightarrow \bar{x}$:

- particle $\rightarrow$ anti-particle
- $\vec{p} \rightarrow -\vec{p}$

$$-2\text{Re}(M(\bar{x})_S^* M(\bar{x})_P)$$



$$-\mathcal{O}_{cp}(\bar{x})$$



Optimal Observable(OO):

$$O_{OO} = \frac{2\text{Re}[\mathcal{M}_{SM}^* \mathcal{M}_{CPV}]}{|\mathcal{M}_{SM}|^2} \quad l_{OO}=144$$

Azimuthal Angle:

$$\mathcal{M}_{\text{int}} \sim \epsilon_{\mu\nu\rho\sigma} p_i^\mu p_j^\nu p_k^\rho p_l^\sigma \rightarrow \vec{p}_a \cdot (\vec{p}_b \times \vec{p}_c)$$

Reconstruction status

- Semileptonic selection: 6% lepton lost, 23% contaminated by other channels
- 50% events exist mis-matched jets
- Nearly random guess on down-type identification

- Pseudo scalar coupling also alters the spin distribution of the top quark, which allow us detect the CP interference from the spin correlation of the top decay products, where **down-type fermion** is recognised as the maximal spin analyser.
- Project partial information of the whole matrix element but easier and clearer, affected less from the reconstruction issue

Machine Learning:

Reconstructed quantity



Truth quantity



Reco-level OO

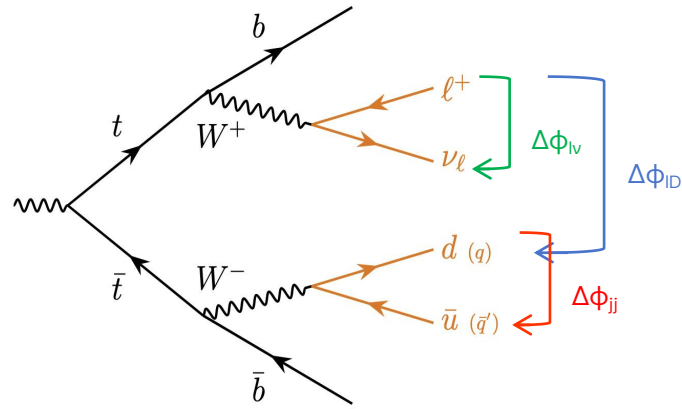
For CPV: Train binary class targeted as $\text{sign}(2\text{Re}(M_S^* M_P)) \rightarrow P(+), P(-)$

$$O_{ML} = P(+) - P(-) \in [-1, 1]$$

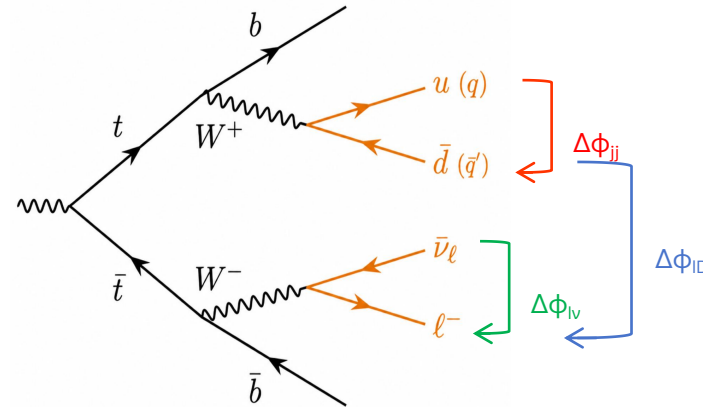
CPV Angular observable baseline at Higgs rest frame

*Higgs rest frame was studied to be the one who contains the most CPV information by Kai Ma in 1809.07127, and we also proved it with our data

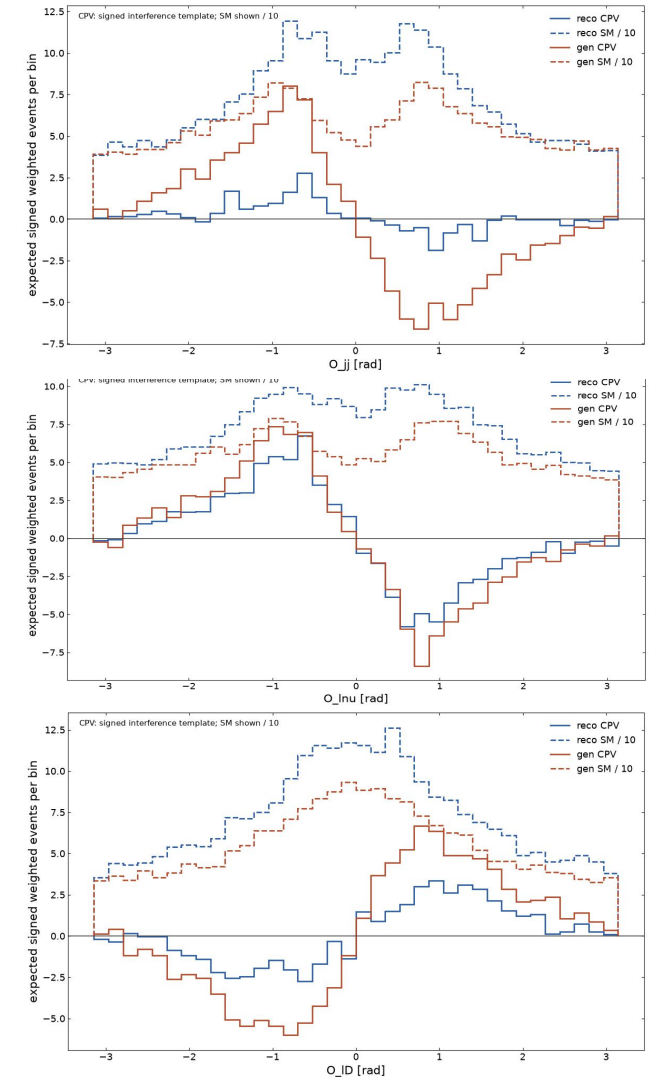
(a) lepton charge > 0



(b) lepton charge < 0

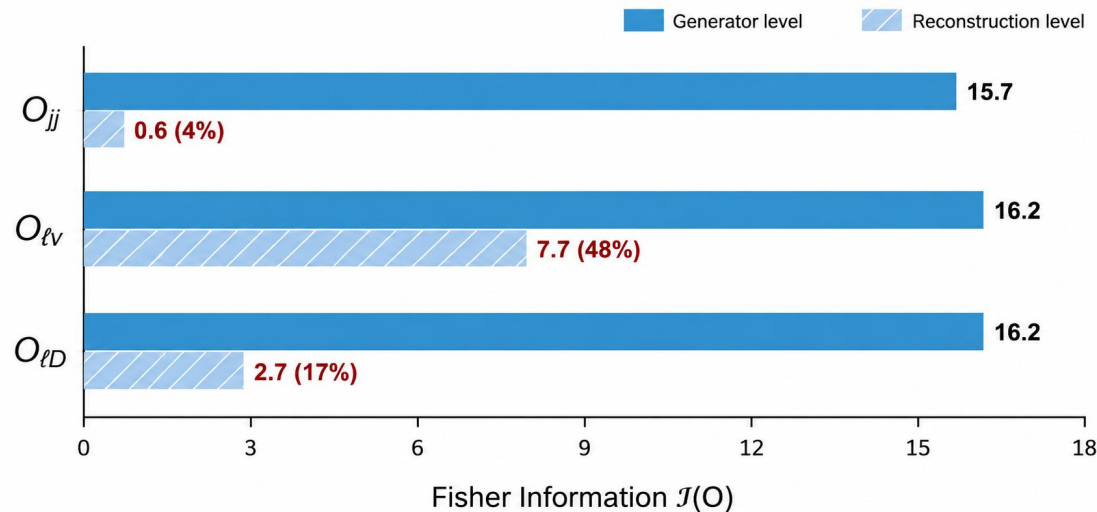


Angular distribution of the muon channel



Define the down-type

Fisher information combined muon and electron channel



- option1: According to the lepton charge

$$P_q(j) = \sum_{f=u,d,s,c} P_f(j),$$

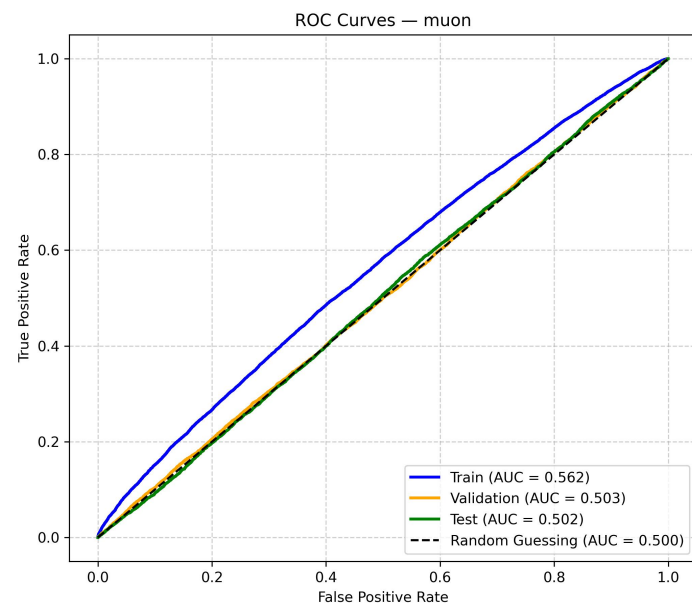
$$P_{\bar{q}}(j) = \sum_{f=u,d,s,c} P_{\bar{f}}(j)$$

- option2: sum over the probability of $s, d, \bar{s}, \bar{d}$

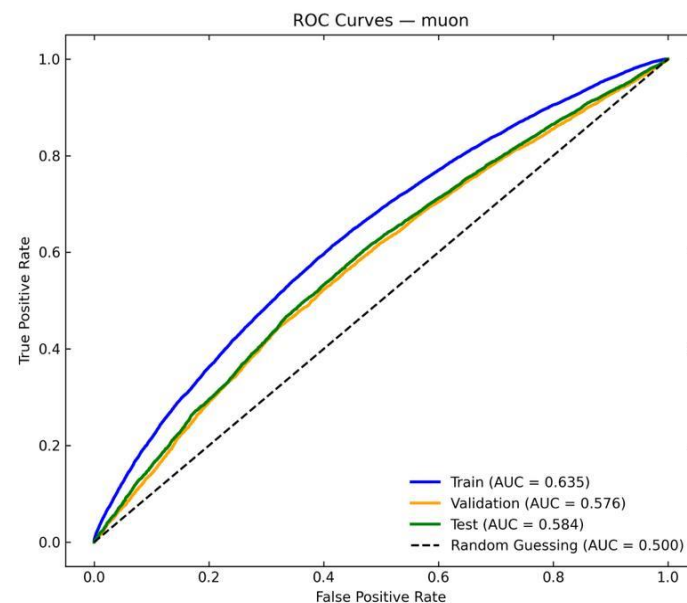
ML observable: From 0 to 1

Angular O_{ID} vs ML[ID]

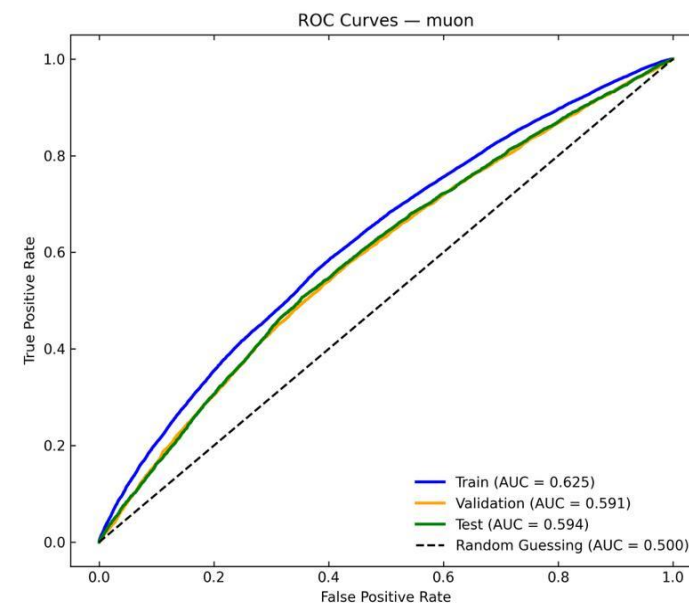
lepton + w down type daughter[E,θ,pT,φ)



+lepton charge



->top side fermion+ anti-top-side fermion



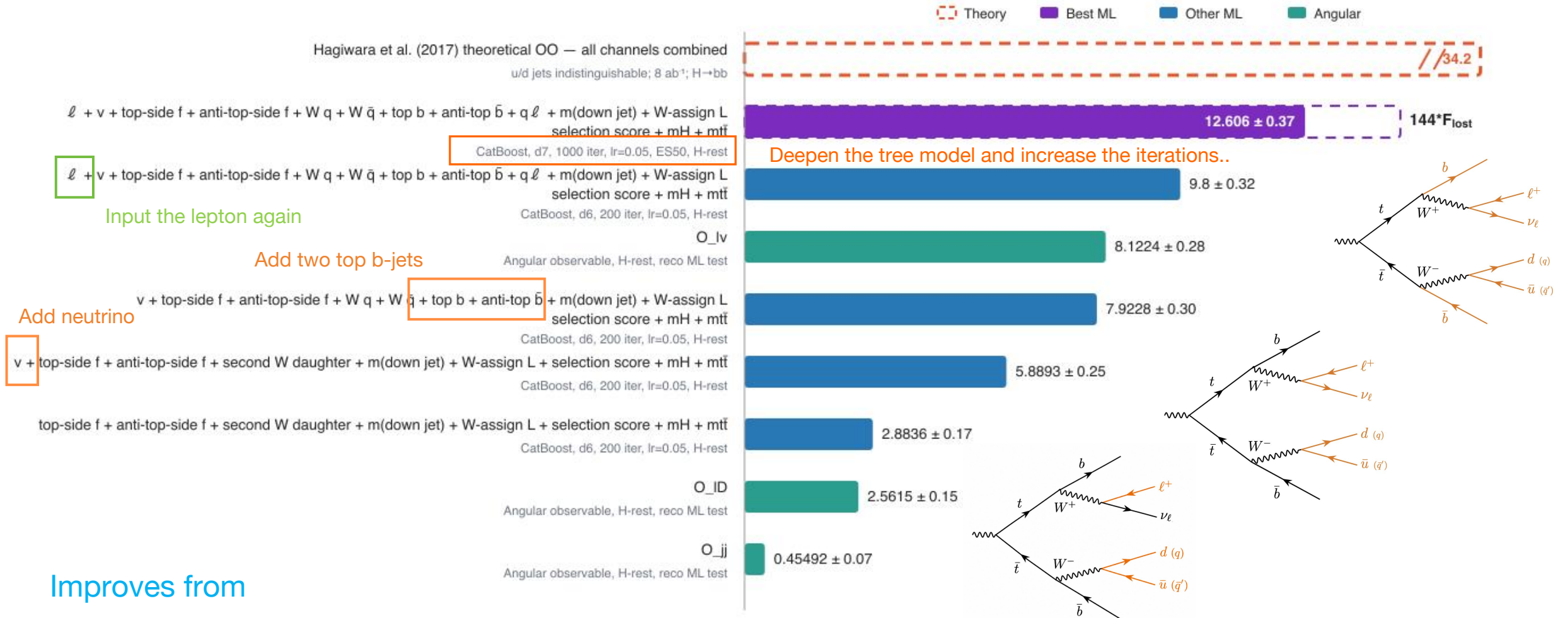
- w assignment likelihood:

$$L_{12} = P_q(j_1)P_{\bar{q}}(j_2),$$

$$L_{21} = P_q(j_2)P_{\bar{q}}(j_1)$$

* much more space can be played around here

ML observable: Approach to the Optimal at Reco Level

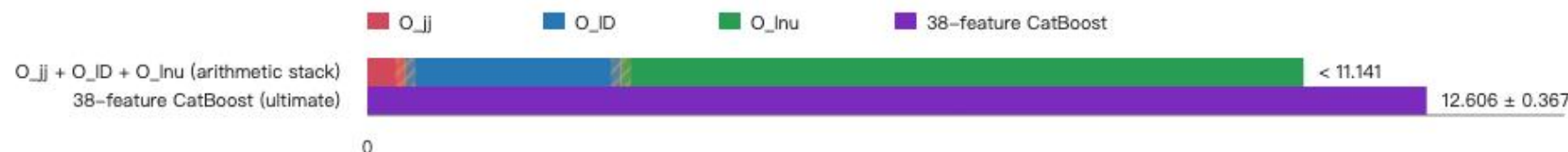


Improves from

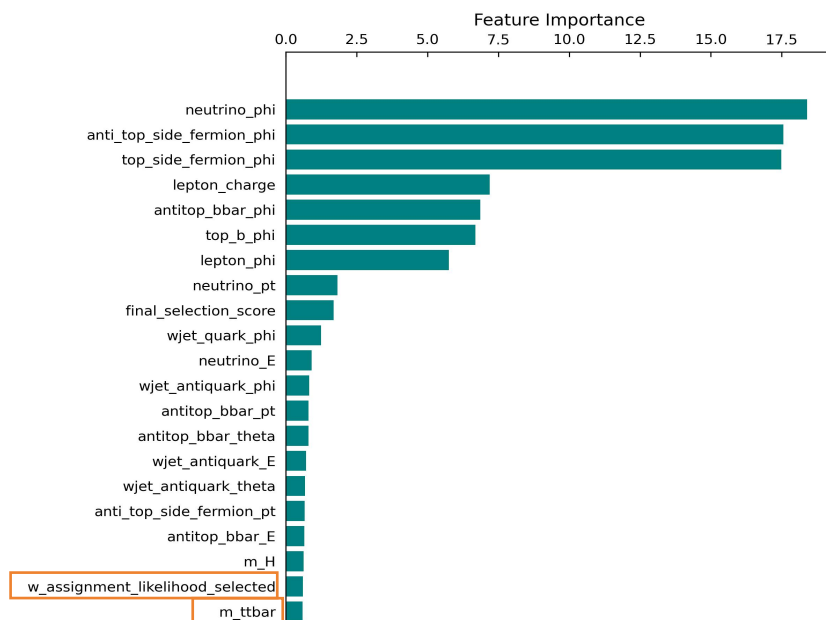
1. Aware the model the order of the input objects
2. Aware the model the reliability of the event reconstruction (jet assignment and the down-type tagging)
3. Input the whole ttbar decay products with both ordered and un-ordered representation
4. Deeper learning

ML observable: What is the model actually learned

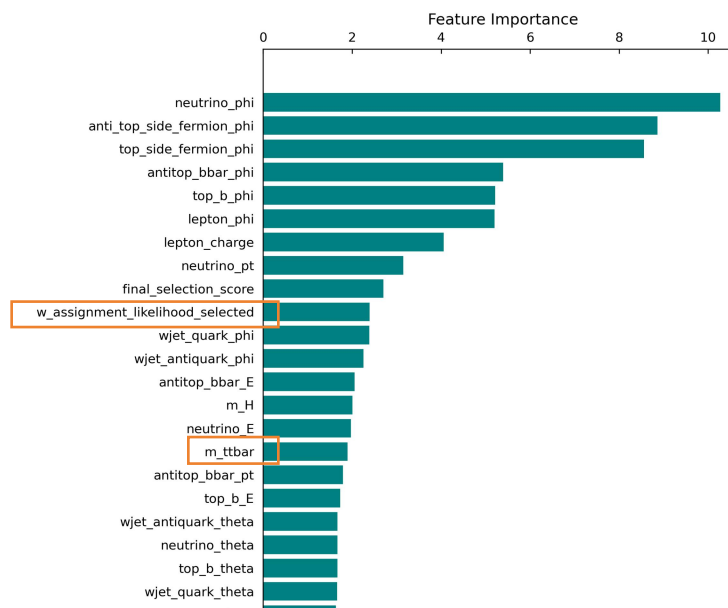
- Project the full spin correlation interference $M_{S,\lambda}^* M_{P,\lambda'}$ into a one-dimensional observable



- Feature importance changing after tuning the model hyper parameters



catboost, ndepth=6, 200iter, learning rate=0.05



catboost, ndepth=7, 1000iter, learning rate=0.05

- Auxiliary variables related to the whole event topology become to gain more importance
- The distribution become more balance

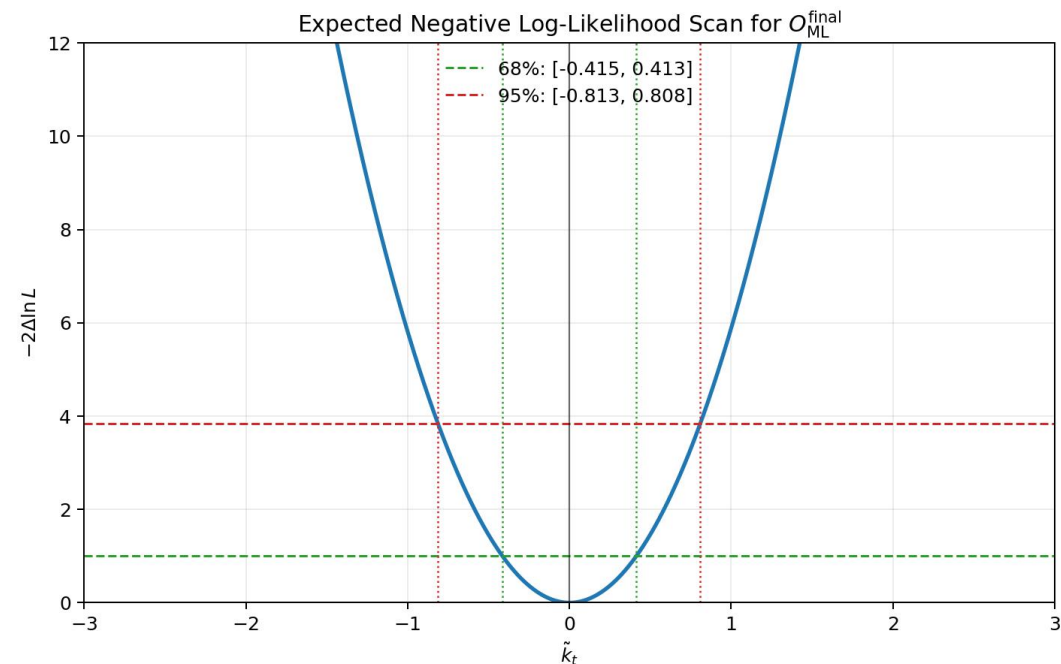
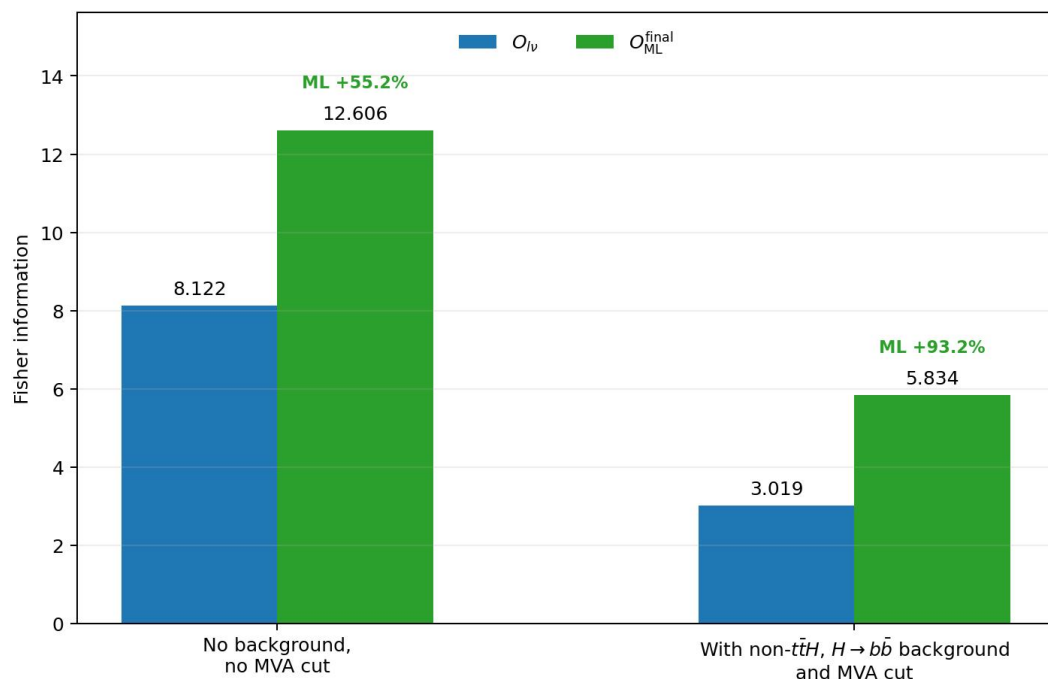
Other kinematics such as θ , p_T show less relevant to the training.

ML observable: A preliminary limit setting with background

Comparison after applying background

Evaluated by $q_{\text{sel}} > 0.954$, 8ab^{-1} , eLpR

Fisher information comparison



Translate to other paramters

$$\mathcal{L}_{t\bar{t}H} = -\frac{m_t}{v} H \bar{t} (\kappa_t + i\tilde{\kappa}_t \gamma_5) t$$

- CP-mixing angle $\kappa_t = \kappa \cos \alpha = 1, \tilde{\kappa}_t = \kappa \sin \alpha$

$\alpha \leq 0.12\pi$ at 68%, $\alpha \leq 0.22\pi$ at 95%,

- SMEFT $\kappa \cos \alpha = 1 - \frac{v^3}{\sqrt{2} m_t} \frac{C_{t\varphi}}{\Lambda^2}, \quad \kappa \sin \alpha = -\frac{v^3}{\sqrt{2} m_t} \frac{C_{t\varphi}^I}{\Lambda^2}.$

$$\left| \frac{C_{t\varphi}^I}{\Lambda^2} \right| \leq 6.7 \text{ TeV}^{-2} \text{ at 68\%} \quad \left| \frac{C_{t\varphi}^I}{\Lambda^2} \right| \leq 13.2 \text{ TeV}^{-2} \text{ at 95\%}$$

ML observable: What can we expect further

Learn the reconstruction ambiguity

- **Fundamental side:**
 - Higher accuracy of light jet tagging from ParT for tth samples
- Optimised the Binary Model :
 - Construct new w-assignment likelihood
 - Inform the b/c-tag
 - Combine the two options as charge/down-type together
 - Add the two b-jets assigned to Higgs, let the model choose it by itself

Learn SM and the background:

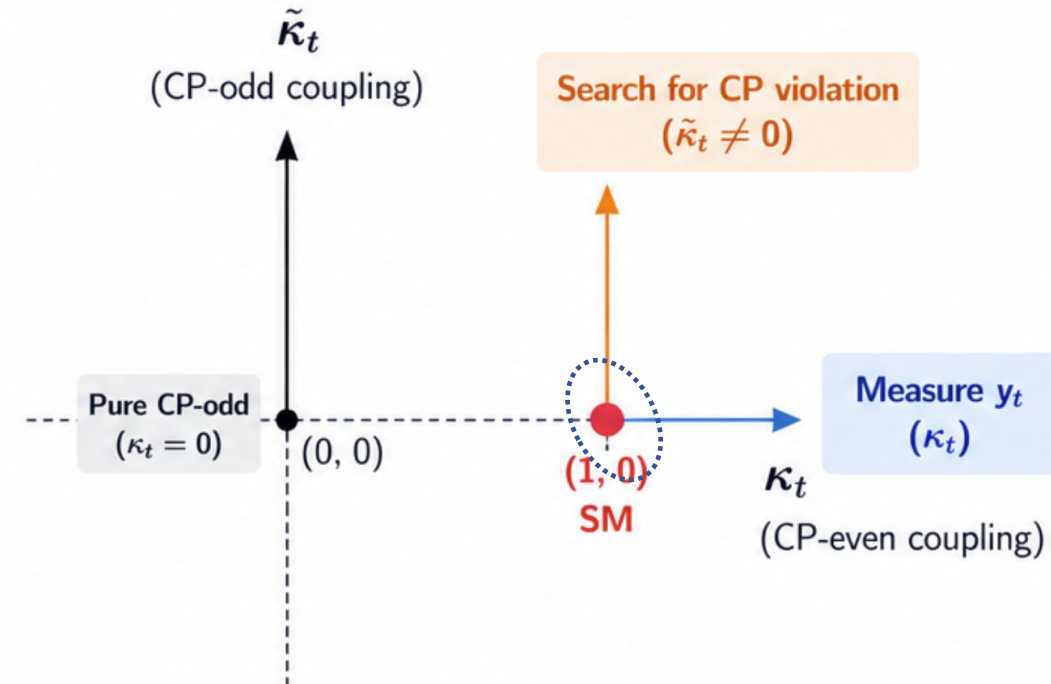
- Train three class model by adding additional SM class
- Add the background into the SM dataset, train them together

Train 4 models to combine the 4 polarisation configuraton at LCF

Summary

- After the implementation of the modern reconstruction tools, the top-yukawa coupling strength can be constrained to accuracy of 2% in semileptonic channel, nearly 100% improves from the 2011 study.
- The ML observable has been studied to exploit the CP information from the CP-odd term directly, the best model so far constraints the pseudo-scalar coupling $|\tilde{\kappa}_t| \leq 0.41$ at 68% CL.
- Many potential method to improve the ML observable to the optimal level in the future.
- The likelihood constructed by kinematic fitting and ParT plays the most important role in this analysis for both side.
- The CP-even term and the CP-odd term will be constrained together later.

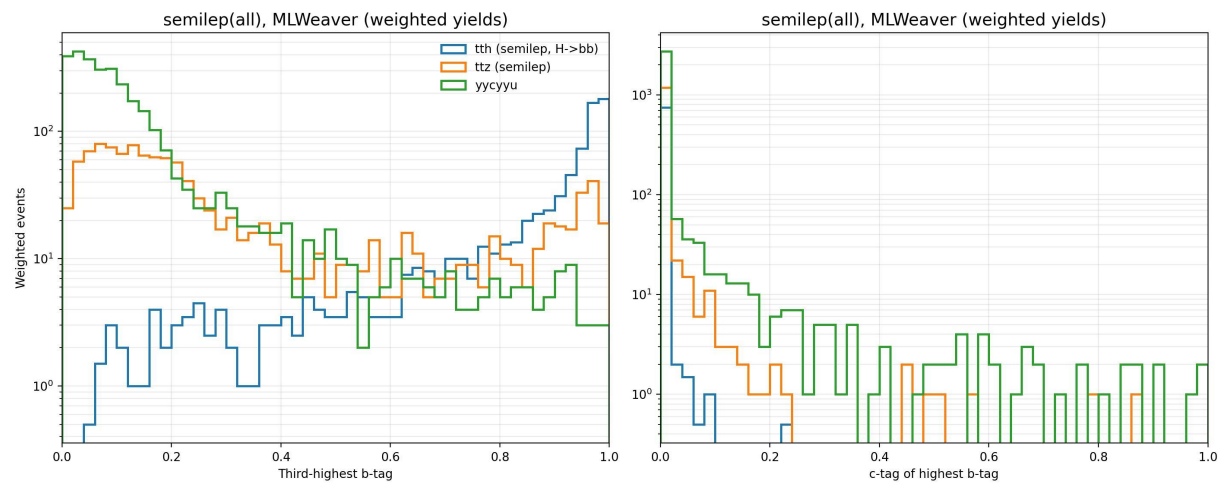
The $(\kappa_t, \tilde{\kappa}_t)$ plane



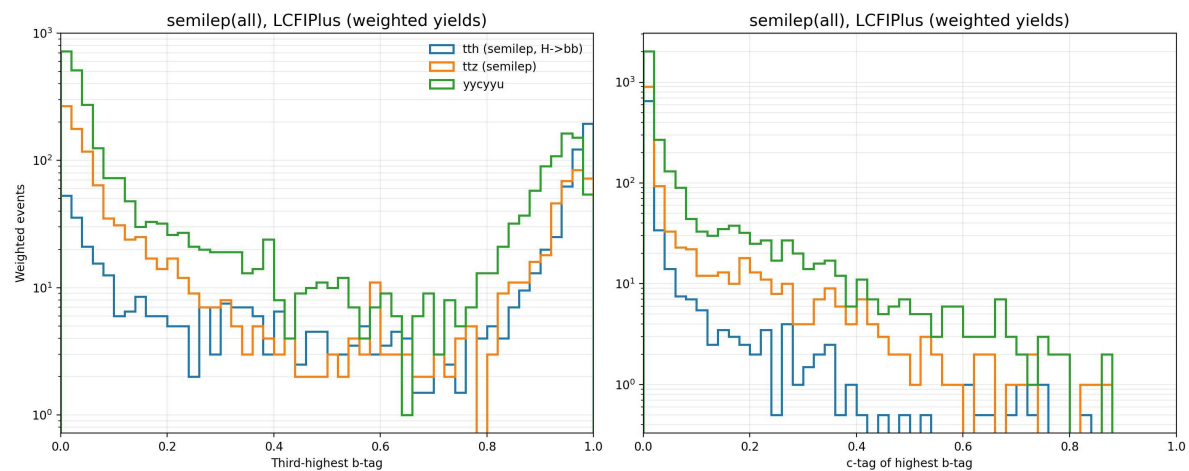
Backup

More Flavor Tagger performance : ParT vs LCFiPlus

ParT

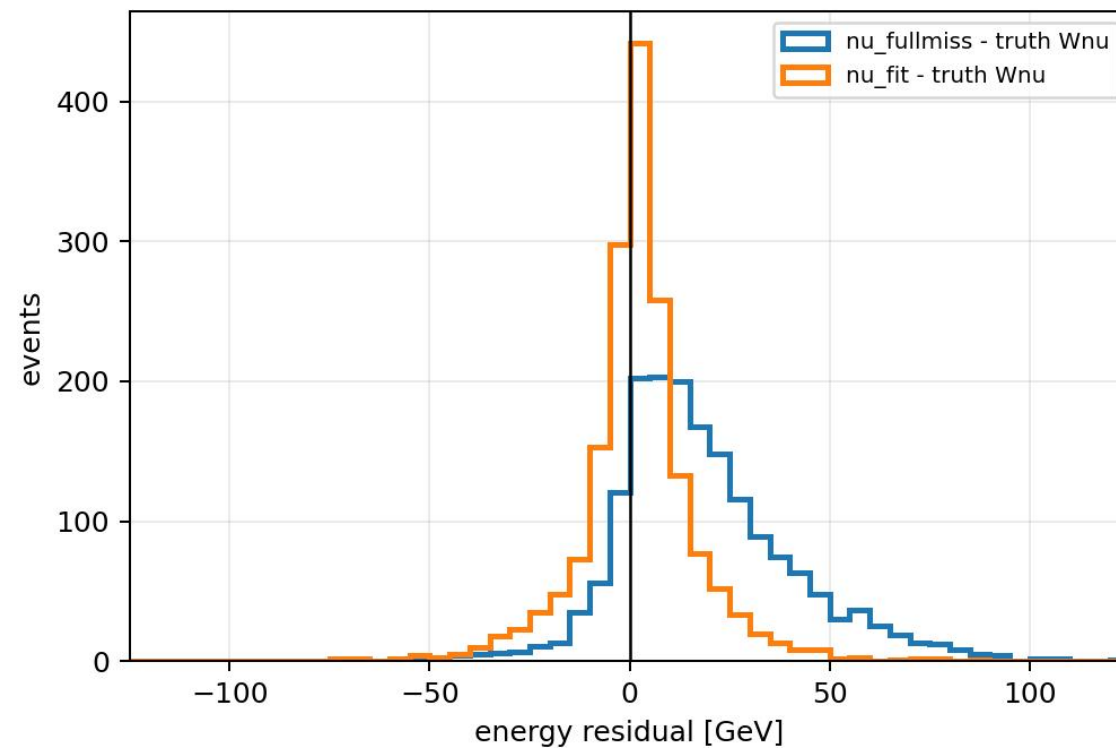
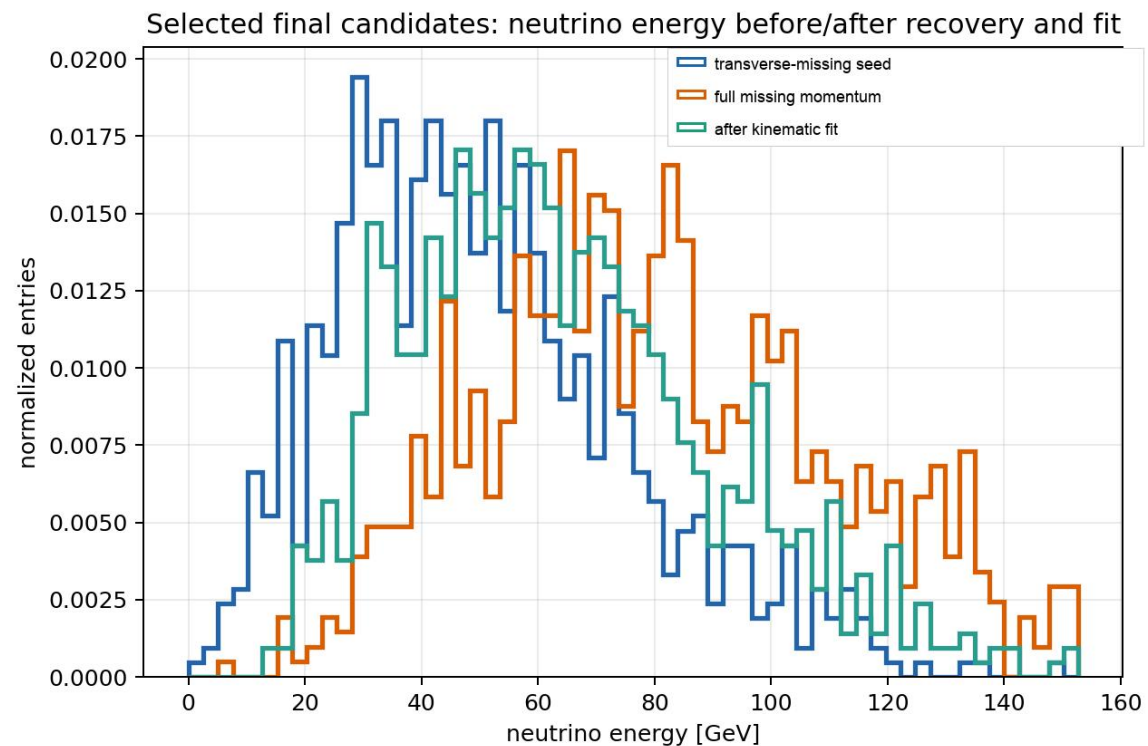


LCFiPlus



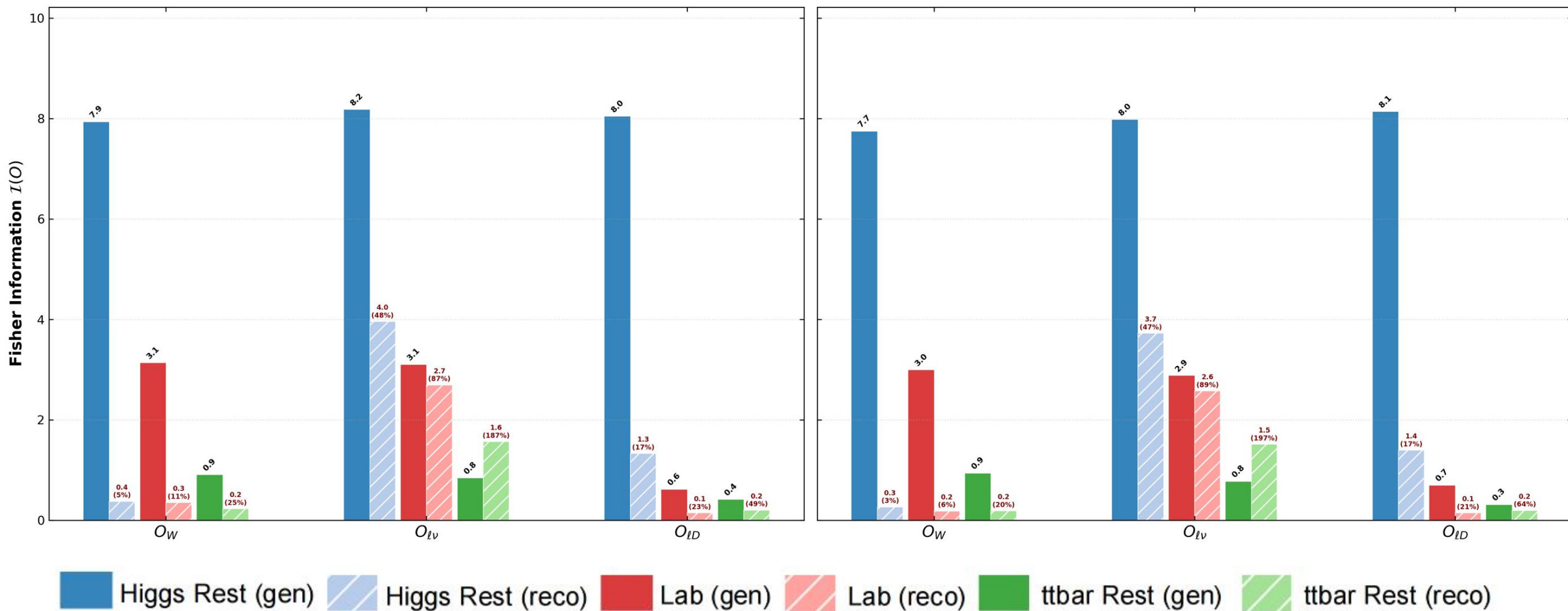
More Flavor Tagger performance and Kinfit details

Kinematic Fitting: Neutrino Correction



Benefit the observable O_{ln}

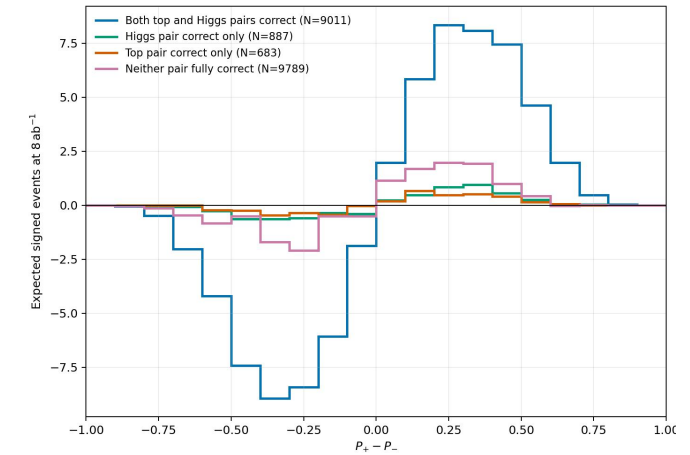
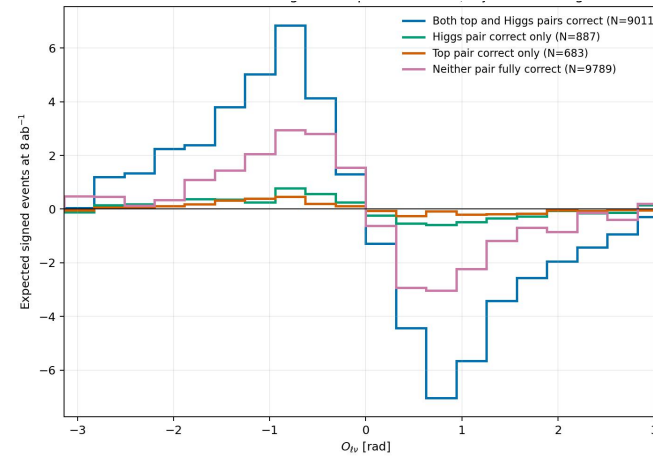
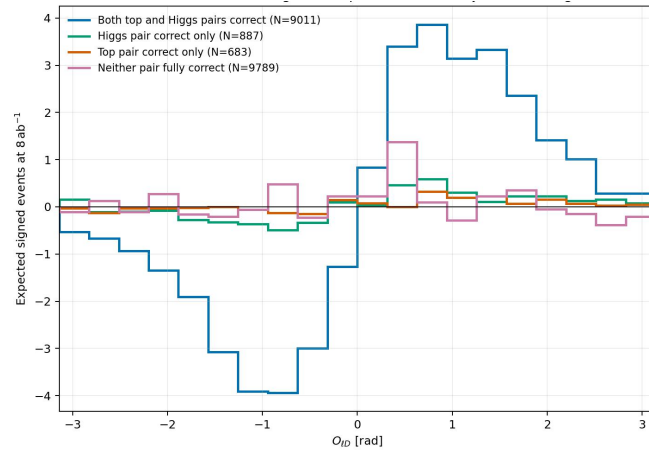
Angular observable baseline in different reference frame



$$\mathcal{M}_P(\lambda_f = s) \propto \begin{cases} \gamma_\psi \beta_\psi & \text{in } R_\psi, \\ \gamma_h \beta_h & \text{in } R_h, \end{cases} \quad \frac{\gamma_h \beta_h}{\gamma_\psi \beta_\psi} = \frac{m_\psi}{m_h} \gtrsim 2.77,$$

$$m_\psi = m_{t\bar{t}} \approx 370 \text{ GeV} \\ m_H \approx 125 \text{ GeV}$$

Comparison of truth-matched jet set

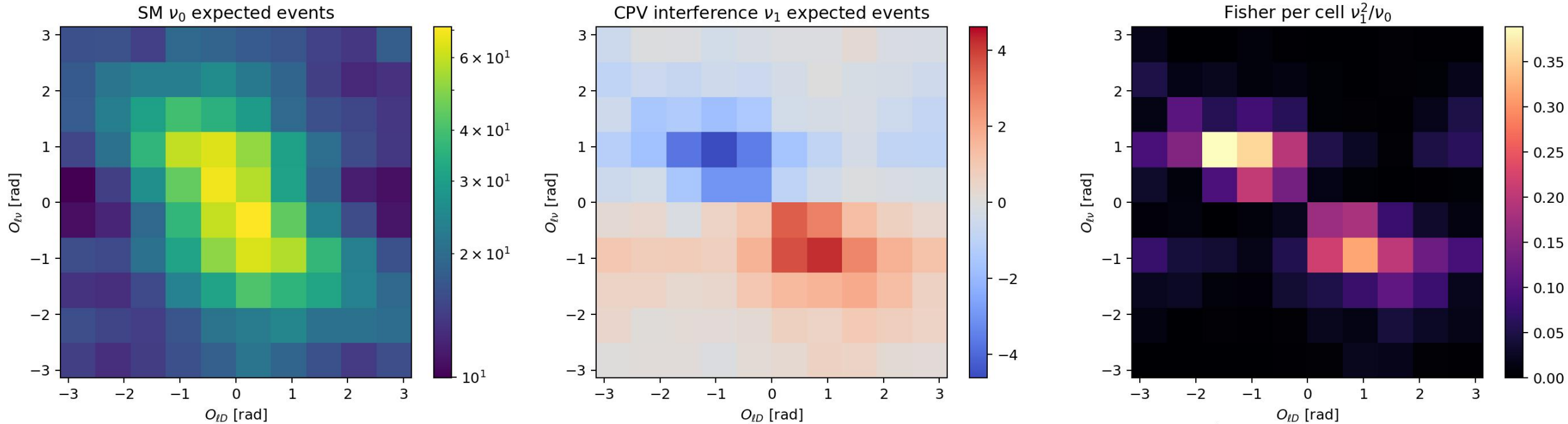


fisher	O_ID	O_Inu	ML_best
Blue	1.93	3.79	7.97
Pink	0.04	0.70	0.27

- The ML fisher value of Blue Line is higher than the value we get by the full event set/2 (half correctly assigned), which actually shows the mis-matched jet events cancel with the correct events

Information fusion of the Angular observables

Reco CPV/SM ML-test 2D angular Fisher: electron, 10x10, $I=4.7971$, invalid=0



This tests whether the joint O_{ID} - O_{IV} template retains more CP interference information than either 1D projection; Higgs-rest reco, test split, 8 ab^{-1} .

$I_{2D_electron}=4.79$, $\mu_{electron}=4.41$ $I_{2D}=9.2$
 $I_{ID}+I_{IV}=10.4 > I_{2D}$