Searching for CP violation with a BDT in ttH multilepton final state with Run 3

Réunion du groupe Particules

Giorgio Mauceri[1][2]

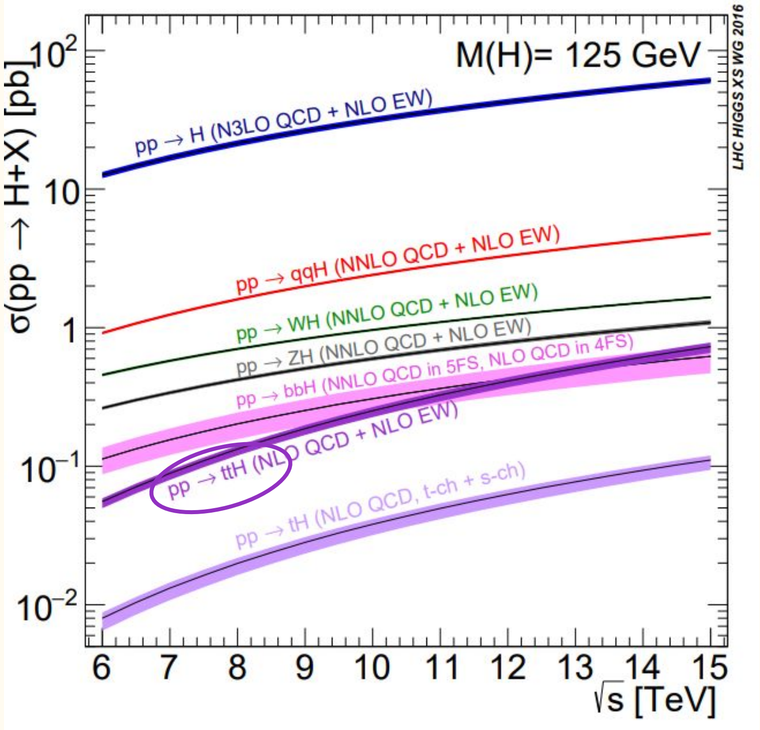
Adriano Di Florio[2], Andrea Giammanco[3],
Jindrich Lidrych[3], Nicolas Chanon[1], Zak Lawrence[3]

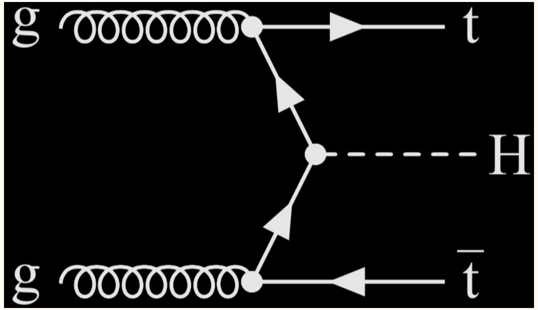
01/12/25

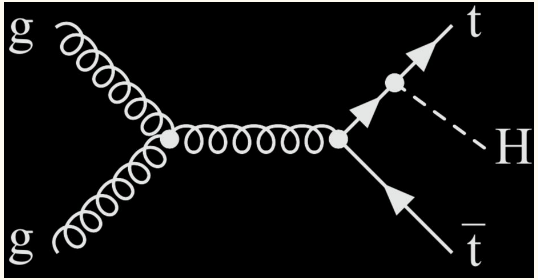
1=IP2I Lyon 2=CC-IN2P3 Lyon 3=CP3 Louvain

Summary

ttH process and CP-violation



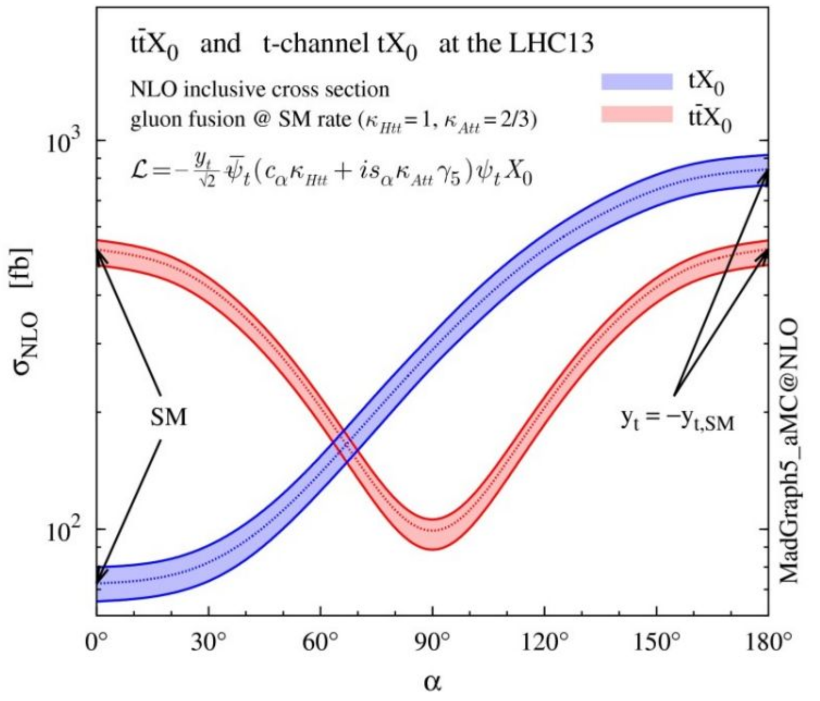




ttH process and CP-violation

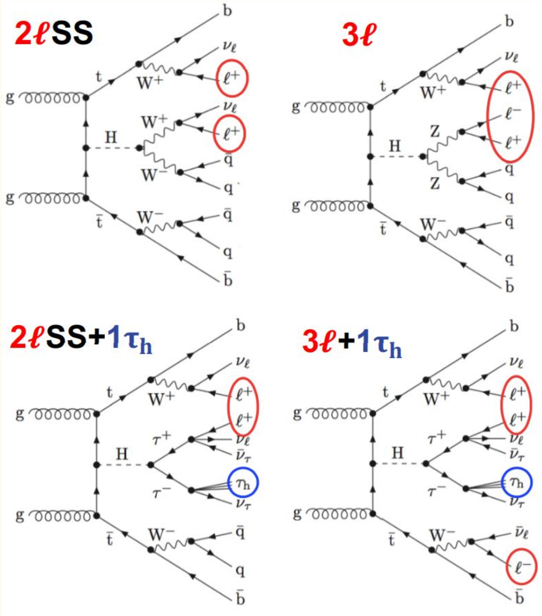
$$\mathcal{L} = -\frac{y_t}{\sqrt{2}}\,\bar{\psi}_t\big(\underbrace{c_\alpha \kappa_{Htt}}_{\text{CP-even}} + \underbrace{i s_\alpha \kappa_{Att}\gamma_5}_{\text{CP-odd}}\big)\psi_t X_0$$

- **α** is the CP mixing angle (0 or 180° in SM)
- $\kappa_{Htt,Att}$ are dimensionless rescaling parameters
- $\mathbf{c_\alpha}$ **and** $\mathbf{s_\alpha}$ are respectively cos(α) and sin(α), meaning the CP-even and CP-odd terms of the interaction
- y_t is the Yukawa coupling constant of the top quark to the Higgs field
- X0 labels a generic spin-0 particle with CP-violating coupling (in this case, the Higgs boson)



CP transformation also affects m_t, p_T, and η

ttH Analysis and usage of the CP-BDT



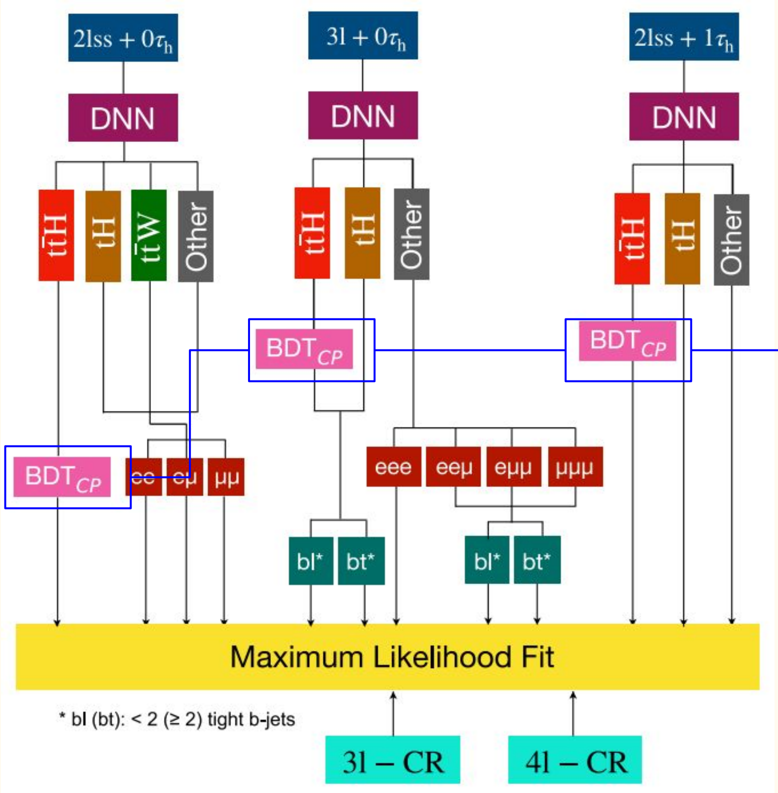


Diagram of the analysis process for ttH

Boosted Decision Trees (BDTs) which discriminate CP-odd from CP-even events in the DNN ttH node based on the score of the CP classifiers

The Dataset used

- The BDT was trained on the TTH CP MC samples (TTH_ctcvcp_4f_TuneCP5_13p6TeV_madgraph-pythia8)

- eras = 2022, 2022EE, 2023, 2023BPix, used all together for the training

- The signal regions analyzed are 2lss0tau and 3l0tau. For now, all events of the signal regions were used, without selecting the ttH node of the multi-target DNN

- The signal was taken as the events with the CP-odd weight, meanwhile the background was taken as the events with the SM weight

- Split into Training and Validation in a ratio 4:1

The Training Method

The BDT is trained using XGBoost, with the following functions:

The evaluation metric is the AUC

Early stopping is enabled. After the training is stopped, the best iteration is recorded.

```
clf = xgb.XGBClassifier(
    tree_method="hist",
    objective="binary:logistic",
    eval_metric="auc", #logloss
    n_estimators=5000,          # early stopping will pick best_n
    subsample=0.8,
    colsample_bytree=0.8,
    learning_rate=0.1,
    max_depth=4,
    min_child_weight=2.0,
    reg_lambda=1.0,
    reg_alpha=0.1,
    gamma=3.,
    random_state=42,
    n_jobs=os.cpu_count(),
    scale_pos_weight=scale_pos_weight,
    early_stopping_rounds=25
)
```

```
clf.fit(
    Xtr, ytr,
    sample_weight=wtr,
    eval_set=[(Xtr, ytr), (Xva, yva)],
    sample_weight_eval_set=[wtr, wva],
    verbose=50
)
```

The Input Variables: Definitions

2lss0tau

Tabella 5: Variables definitions 2lss0τ

Variable Name	Definition
SelJet_pt	pT of leading jet
SelJet_Eta	η of leading jet
SelJet_Phi	ϕ of leading jet
SelJet_Mass	Mass of leading jet
SelJet_isBtag	Btag class of the leading jet
SelJet_isFromHadTop	Whether the leading jet comes from the hadronic top
SelJet_BTagDeepFlavB	Deep flavour Btag of the leading jet
mindRlep1jet	dR of lep 1 to its closest jet
mindRlep2jet	dR of lep 2 to its closest jet
mTTH	invariant mass of jets+met+leptons
dEtaBB	dEta of two jets with highest b tagging score
dEtaLL_BBframe	dη of the two leptons in the B-B system frame
avg_dr_jet	average dR distance among all jets
dEtaBB_LLframe	dEta BB in the l-l system frame
Hj_tagger_hadTop	Higgs-jet tagger
HTT_score	highest BDT score of jet triplet from t
met_phi	ϕ of met
nSelJets	number of jets passing the cuts
met	missing transverse energy

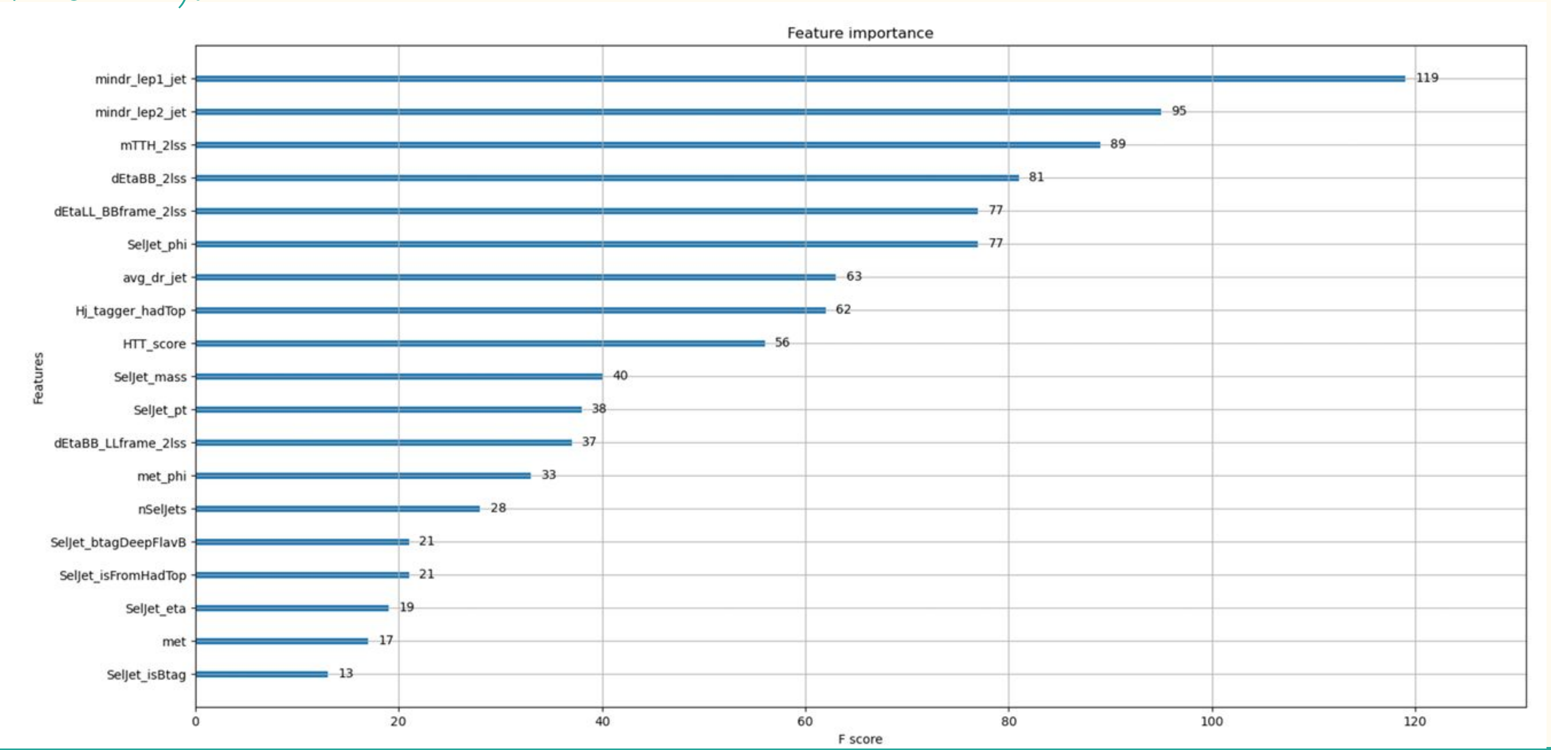
3l0tau

Tabella 4: Variables definitions 3l0τ

Variable Name	Definition
jetpt1	pT of leading jet
jetEta1	η of leading jet
jetPhi1	ϕ of leading jet
jetMass1	Mass of leading jet
jetpt2	pT of subleading jet
jetEta2	η of subleading jet
jetPhi2	ϕ of subleading jet
jetMass2	Mass of subleading jet
Lep1_pT	pT of lepton 1
Lep2_pT	pT of lepton 2
Lep3_pT	pT of lepton 3
mindRlep1jet	dR of lep 1 to its closest jet
mindRlep2jet	dR of lep 2 to its closest jet
mTTH	invariant mass of jets+met+leptons
dEtaBB	dEta of two jets with highest b tagging score
dEtaL1L3_BBframe	dη of leptons 1 and 3 in the B-B system frame
dEtaL1L2_BBframe	dη of leptons 1 and 2 in the B-B system frame
dRlep12	dR of lepton 1 and 2
dRlep23	dR of lepton 2 and 3
dRlep31	dR of lepton 3 and 1

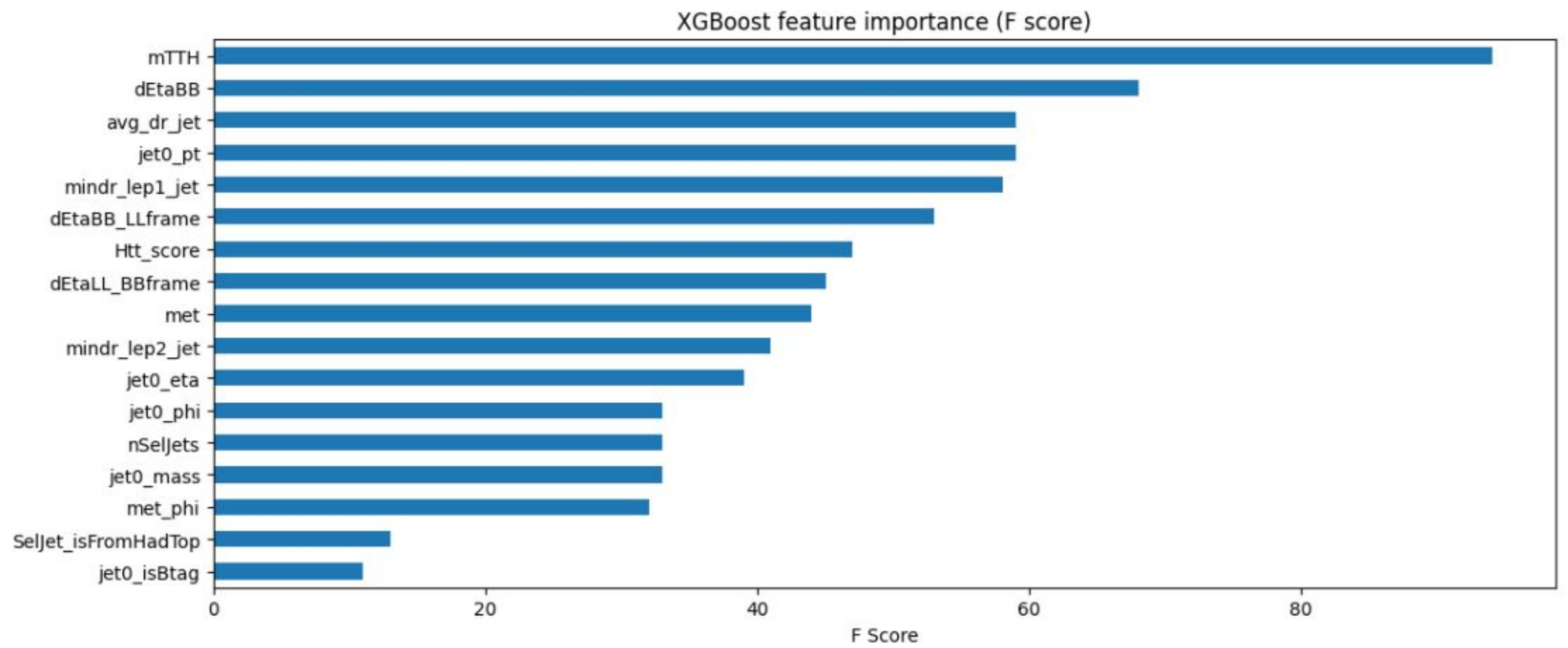
The Input Variables: Variable ranking in Run 2 (2lss0tau)

All features used for the 2lss0tau CP-BDT, with relative importance (from the CMS AN-20-241):



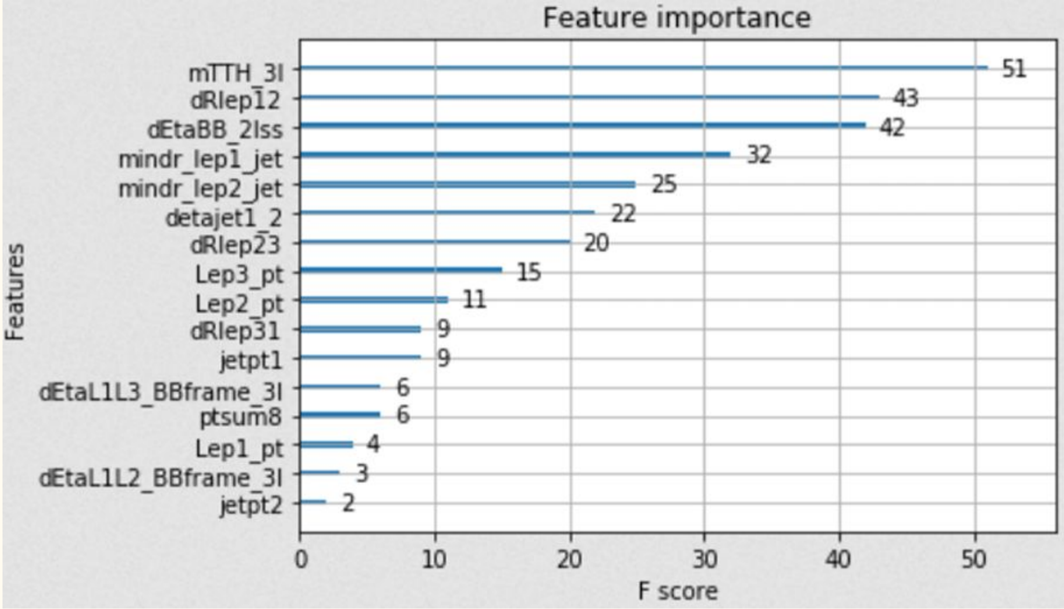
My Input Variables: Variable ranking (2lss0tau)

All features used for the 2lss0tau CP-BDT, with relative importance. The missing features are those relying on the Higgs-Jet tagger, and the Seljet_btagDeepFlavB



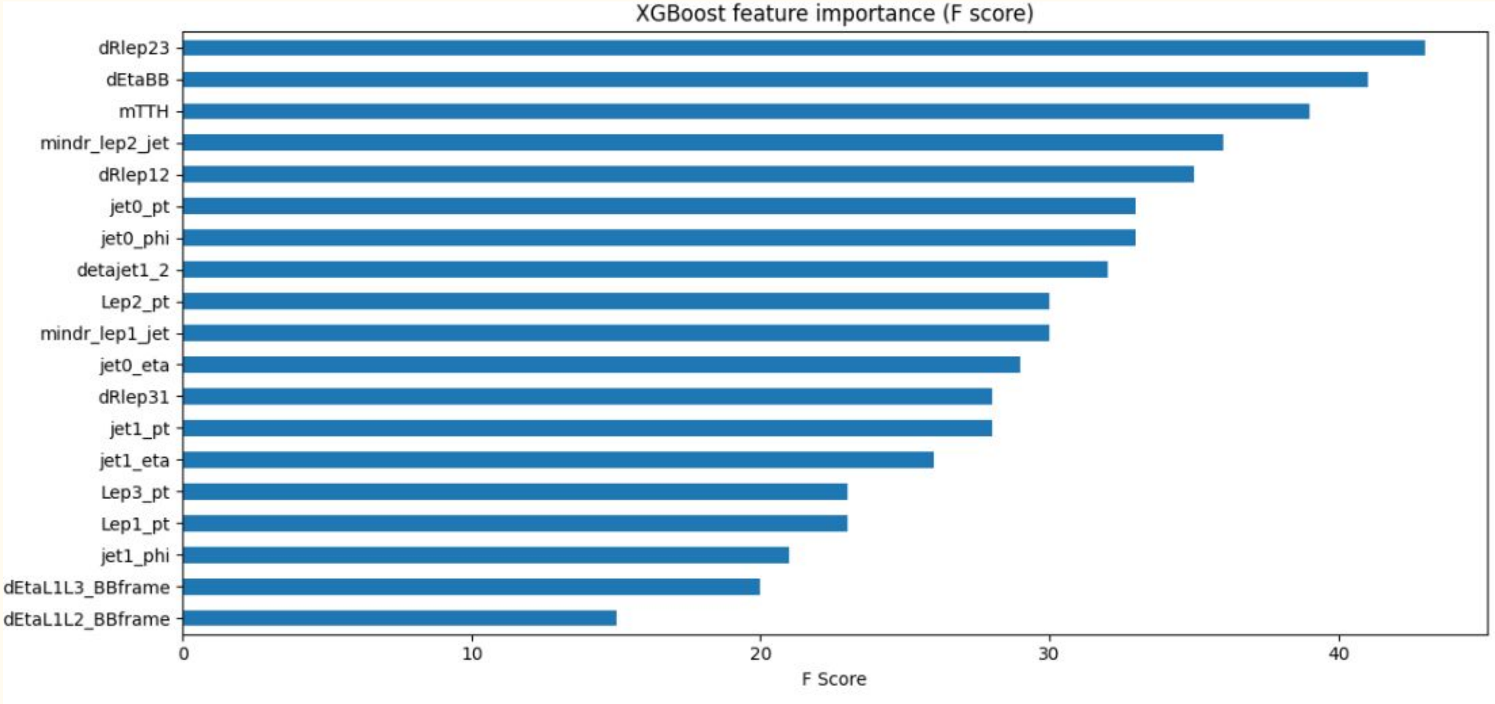
The Input Variables: Variable ranking in Run 2 (3l0tau)

All features used for the 3l0tau CP-BDT in Run2, with relative importance (from the CMS AN-20-241):



My Input Variables: Variable ranking (3l0tau)

Features used for the 3l0tau BDT, with relative importance. All variables from Run2 were used, and extra variables regarding the properties of the jets were added



The Hyperparameters: Ranges and values used in Run 2

The hyperparameters used in the Run 2 analysis (from the CMS AN-20-241):

Table 10: Range of tested hyperparameters

Hyperparameter	Range	Explanation
learning_rate	[0.01,4]	the rate at which the algorithm learns
n_estimators	[100,1000]	the number of estimators (trees) used
max_depth	[3,6]	the depth of each tree (max. number of features per tree)
subsample	[0.8,1]	the amount of examples used to build each tree
colsample_bytree	[0.8,1]	the amount of features used to build each tree
gamma	0,1,5	a regularization parameter (either 0,1 or 5)
early_stopping	True,False	stops adding new trees if val. loss stops decreasing

Table 11: Optimal choice of BDT hyperparameters

Hyperparameter	$2\ell ss + 0\tau_h$	$2\ell ss + 1\tau$	$3\ell ss + 0\tau_h$
learning_rate (=eta)	0.1	0.05	4
n_estimators	120	120	200
max_depth	4	4	2
subsample	0.8	0.8	1
colsample_bytree	1	1	1
gamma	1	5	0
early_stopping	True	False	True

The Hyperparameters: Ranges used for retraining

The hyperparameter configurations from Run 2 were tried. After some further work, I made the following modifications:

1. always used early_stopping, which in turn made a large number of estimators redundant
2. gamma = 5 was removed, as it gave overall worse results, and often different gamma values give the same output
3. the learning rate was capped at 2.5

Tabella 1: HyperParameters

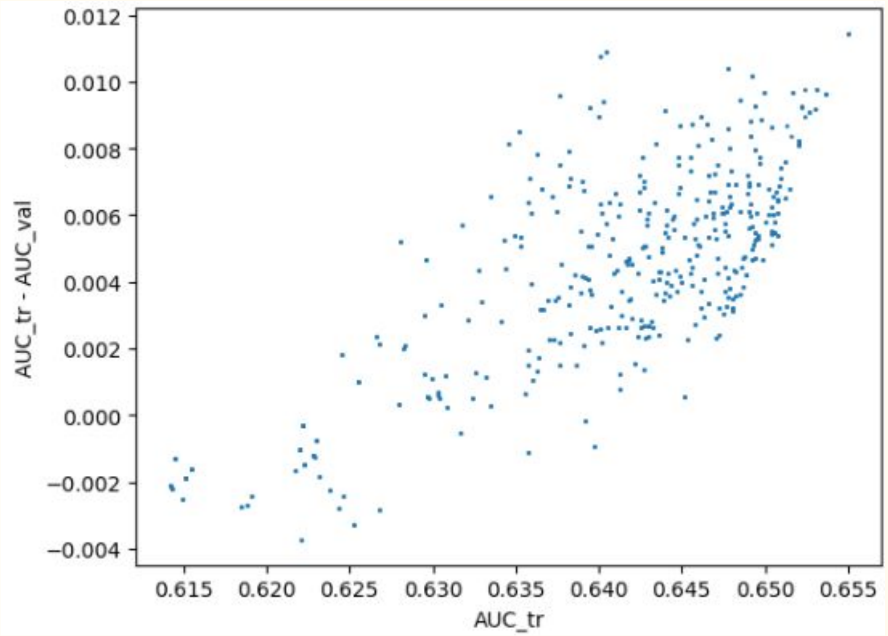
Hyperparameter	Range	Explanation
learning_rate	[0.01, 2.5]	the rate at which the algorithm learns
n_estimators	[1000]	the number of estimators (trees) used
max_depth	[3, 6]	the depth of each tree (max. number of features per tree)
subsample	[0.8, 1]	the amount of examples used to build each tree
colsample_bytree	[0.8, 1]	the amount of features used to build each tree
gamma	[0, 1]	a regularization parameter (either 0,1 or 5)
early_stopping	[True]	stops adding new trees if val. loss stops decreasing

Fine Tuning: Best model choice (2lss0tau)

Since each BDT took only a few seconds to train, a grid search was used to look for the best combinations of hyperparameters, by running over hundreds of possible combinations.

Afterwards the following plot was made. On the x-axis, the AUC for the training set for all hyperparameter combinations used. On the y axis, the difference between the AUC of the training and validation sets.

The hyperparameter combination taken was the one that gave the point that allowed to maximize the AUC while minimizing the difference between the AUC of the two sets:



Fine Tuning: Best model choice (2lss0tau)

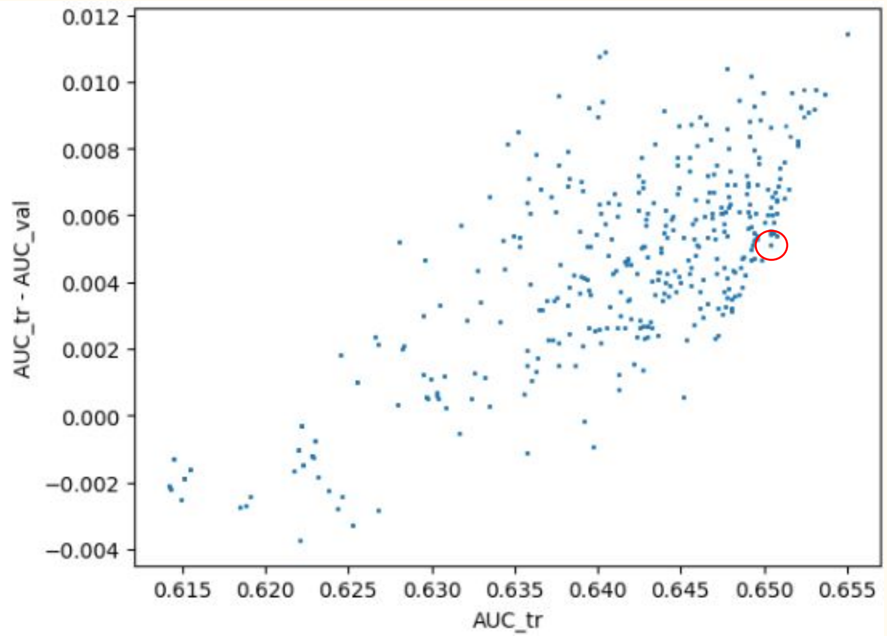
Since each BDT took only a few seconds to train, a grid search was used to look for the best combinations of hyperparameters, by running over hundreds of possible combinations.

Afterwards the following plot was made. On the x-axis, the AUC for the training set for all hyperparameter combinations used. On the y axis, the difference between the AUC of the training and validation sets.

The hyperparameter combination taken was the one that gave the point that allowed to maximize the AUC while minimizing the difference between the AUC of the two sets:

Tabella 2: HyperParameters 2lss0τ

Hyperparameter	Range
learning_rate	0.26
n_estimators	1000
max_depth	3
subsample	0.95
colsample_bytree	1.0
gamma	1
early_stopping	True

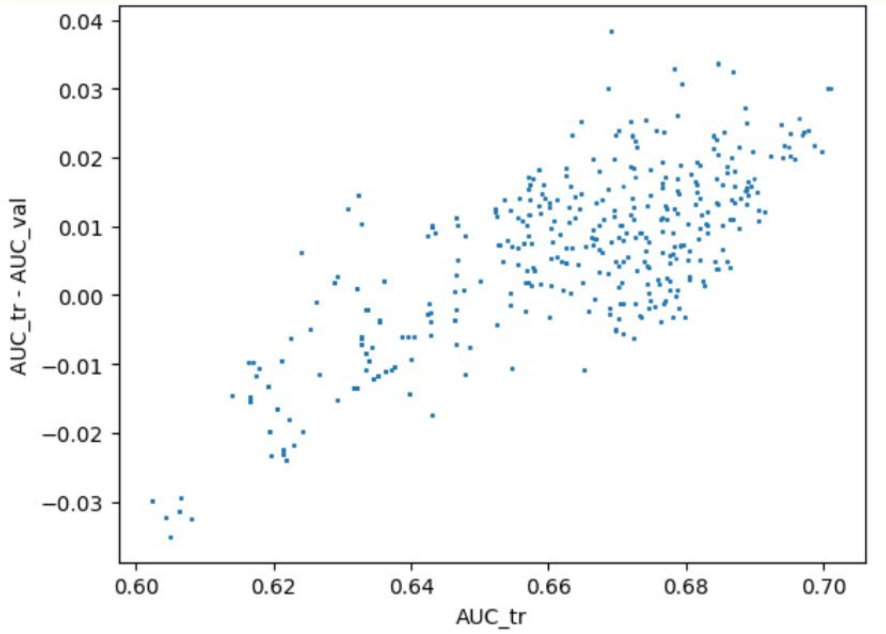


Fine Tuning: Best model choice (3l0tau)

Since each BDT took only a few seconds to train, a grid search was used to look for the best combinations of hyperparameters, by running over hundreds of possible combinations.

Afterwards the following plot was made. On the x-axis, the AUC for the training set for all hyperparameter combinations used. On the y axis, the difference between the AUC of the training and validation sets.

The hyperparameter combination taken was the one that gave the point that allowed to maximize the AUC while minimizing the difference between the AUC of the two sets:



Fine Tuning: Best model choice (3l0tau)

Since each BDT took only a few seconds to train, a grid search was used to look for the best combinations of hyperparameters, by running over hundreds of possible combinations.

Afterwards the following plot was made. On the x-axis, the AUC for the training set for all hyperparameter combinations used. On the y axis, the difference between the AUC of the training and validation sets.

The hyperparameter combination taken was the one that gave the point that allowed to maximize the AUC while minimizing the difference between the AUC of the two sets:

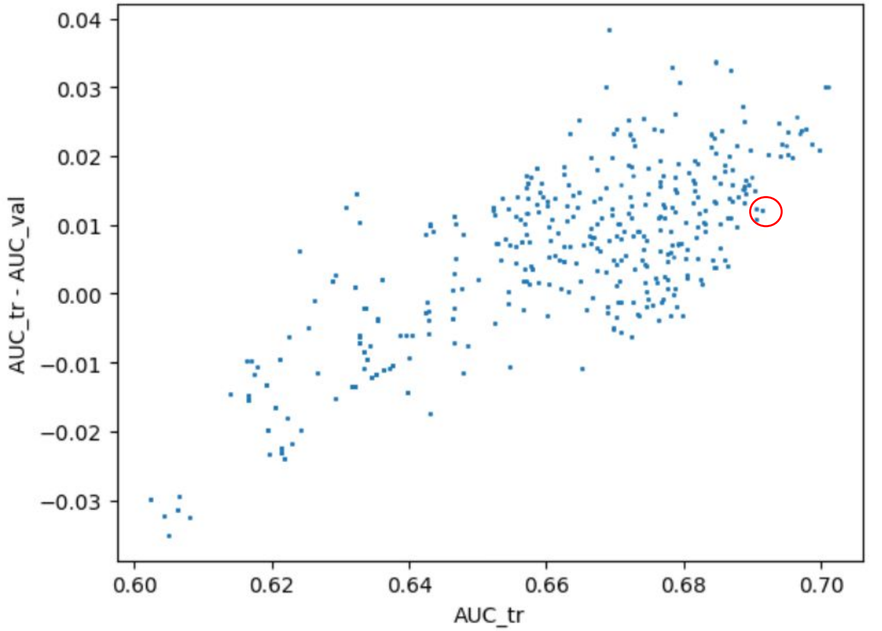
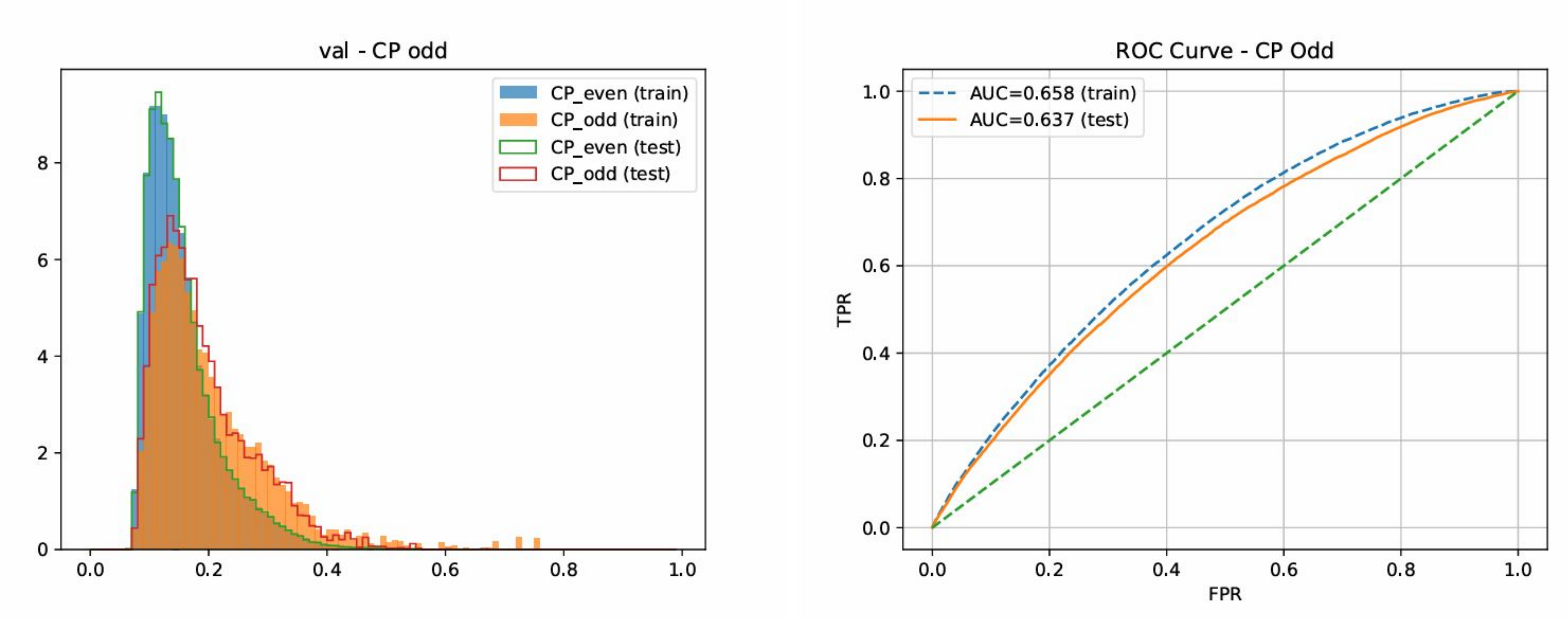


Tabella 3: HyperParameters 3l0τ

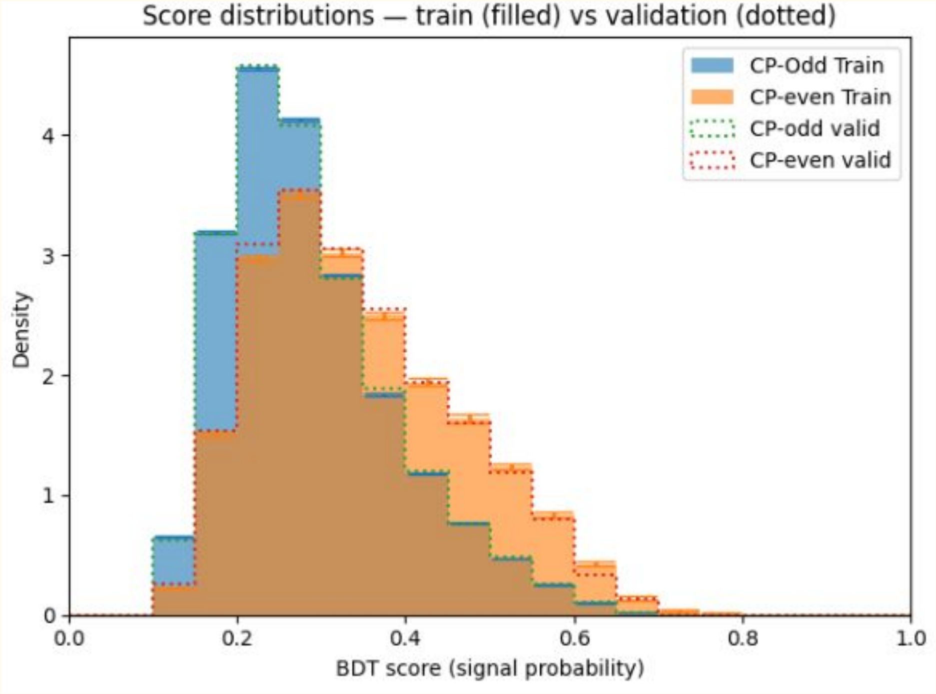
Hyperparameter	Best value
learning_rate	0.26
n_estimators	1000
max_depth	4
subsample	0.9
colsample_bytree	0.8
gamma	1
early_stopping	True

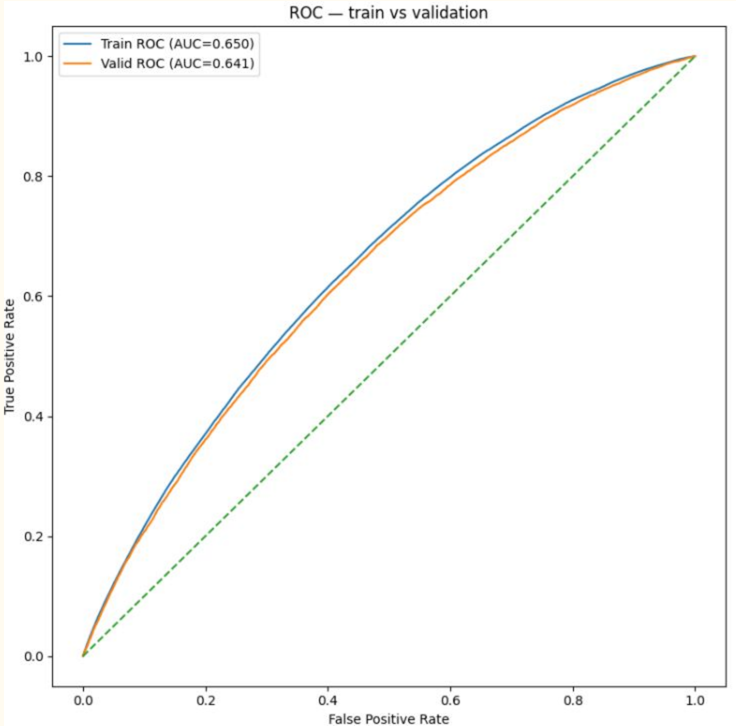
The Output: score and ROC curve in Run 2 (2lss0tau)



BDT score distributions for 2lss0tau for CP-even and CP-odd (left) and corresponding ROC curve with AUC=0.637 (right)

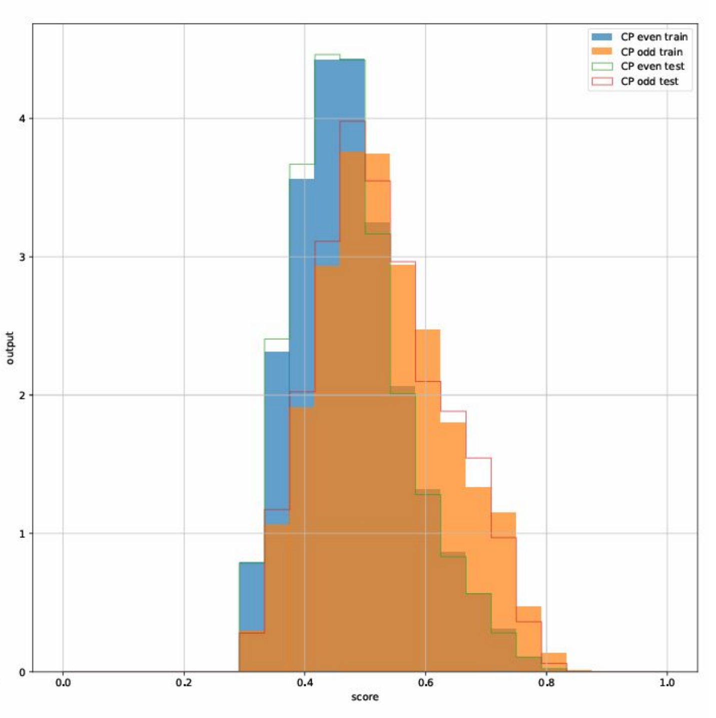
My Output: score and ROC curve (2lss0tau)

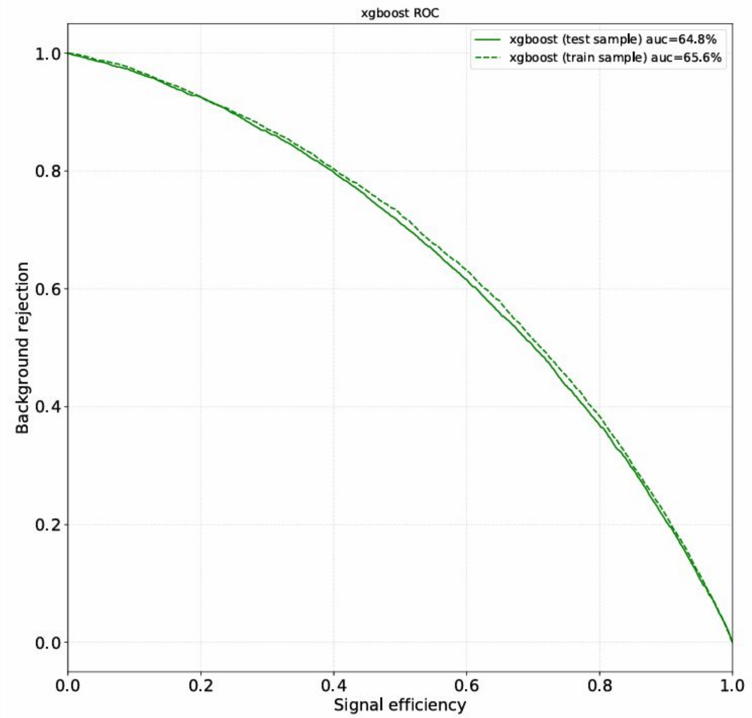




Predicted distributions for 2lss0tau for CP-even and CP-odd (left) and corresponding ROC curve with AUC=0.661 (right)

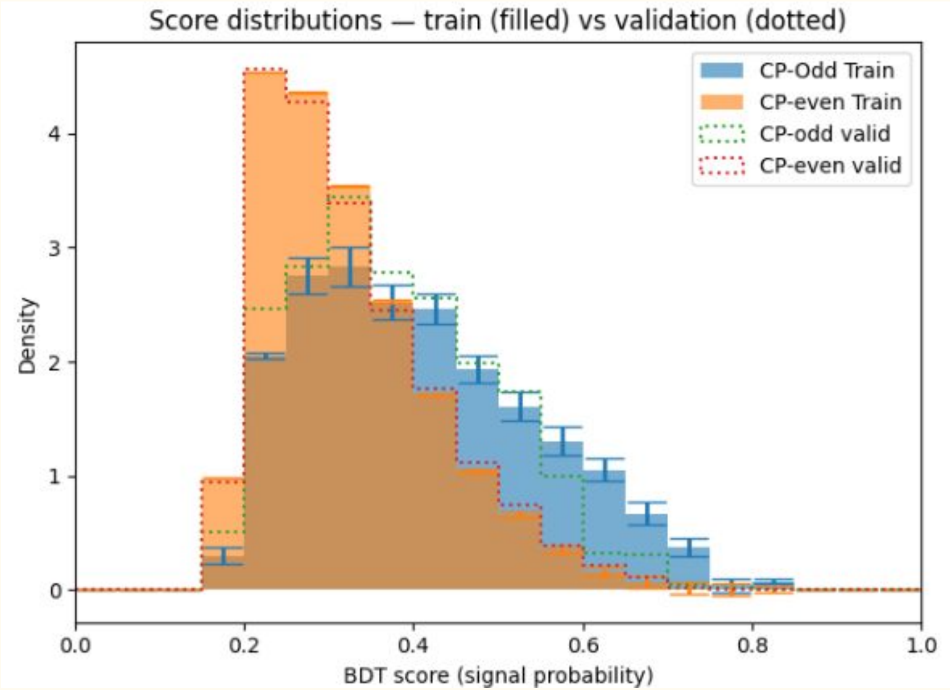
The Output: score and ROC curve in Run 2 (3l0tau)

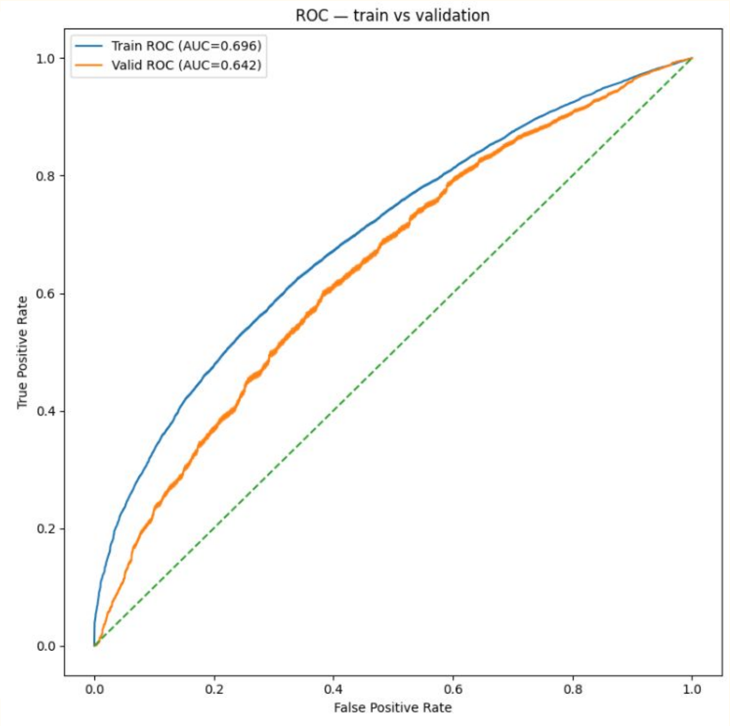




ROC curve for 3l0tau for CP-even and CP-odd with AUC = 0.648 (right) and corresponding predicted distribution (left)

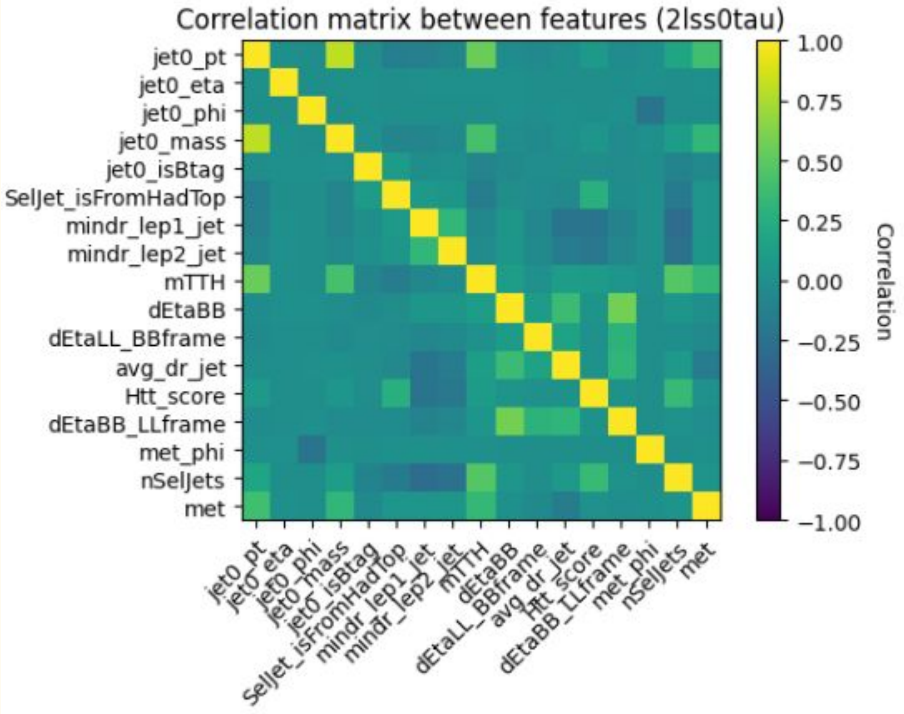
My Output: score and ROC curve (3l0tau)

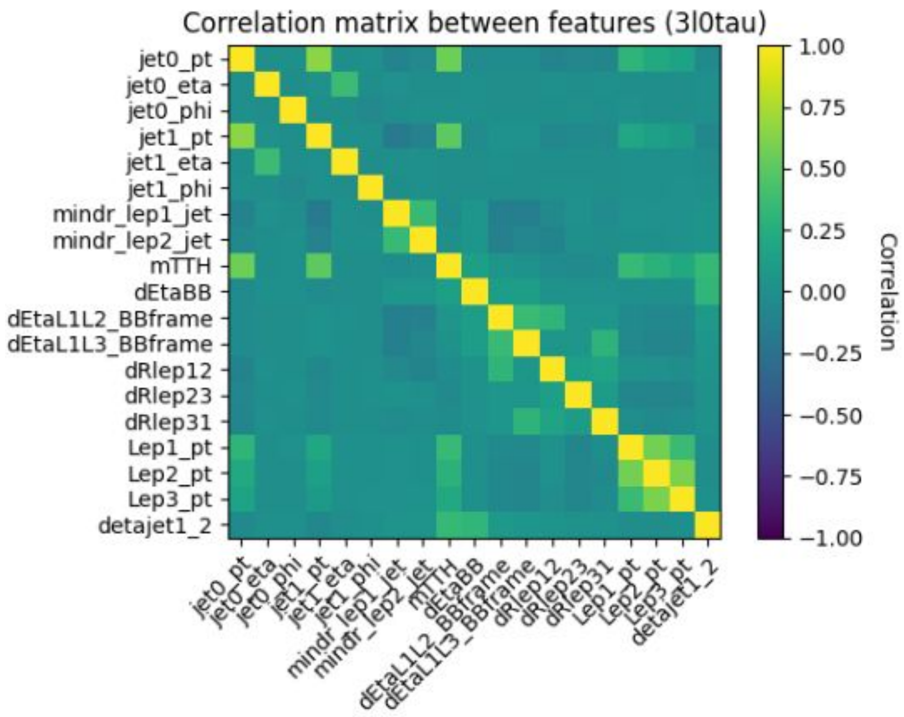




ROC curve for 3l0tau for CP-even and CP-odd (right) and corresponding predicted distribution with AUC = 0.696 (left)

The Output: Correlation matrices in Run3





Next steps:

- Try new variables suggested for CP-sensitive observables in the STXS formalism (from arxiv:2406.03950)

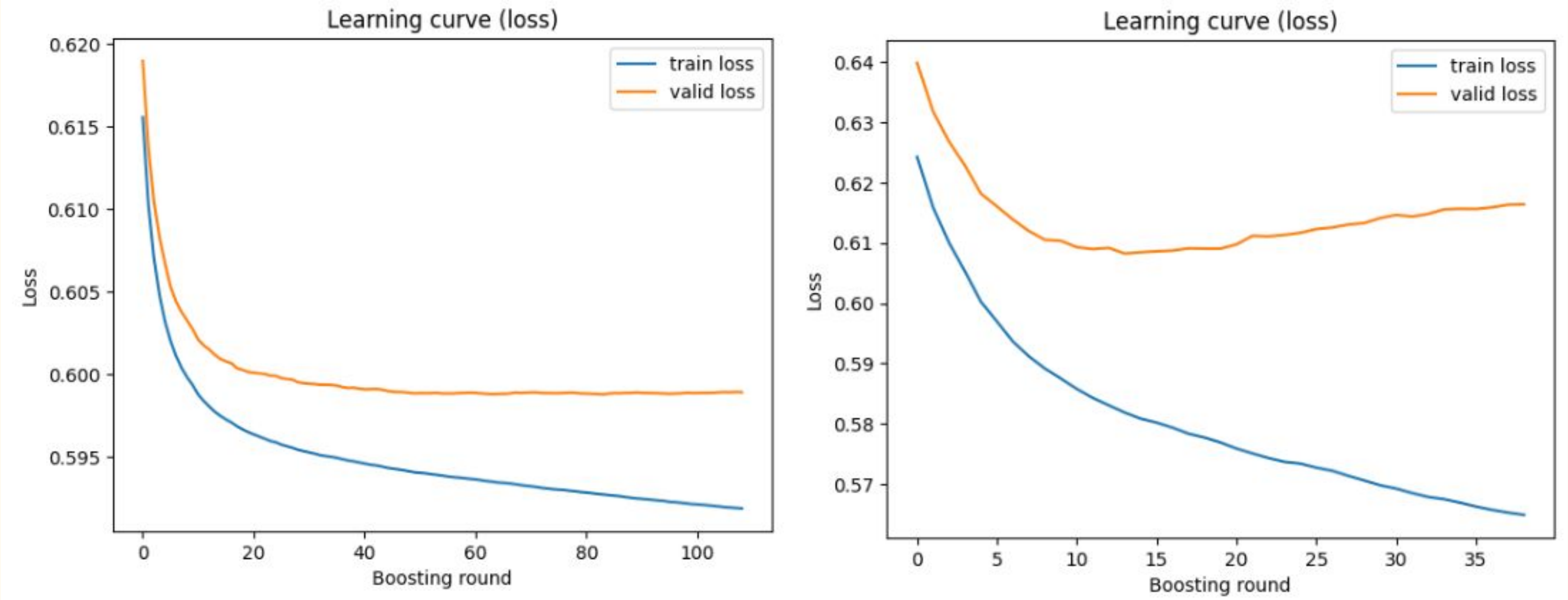
- Apply (and maybe train) the BDT in the ttH node of the DNNs

- Implement the postmortem reweighting for the tHq and tHW samples, and retrain the BDT on those channels too

- Add missing variables from the AN
    - Higgs-Jet tagger needs synchronization

Thank you for the Attention

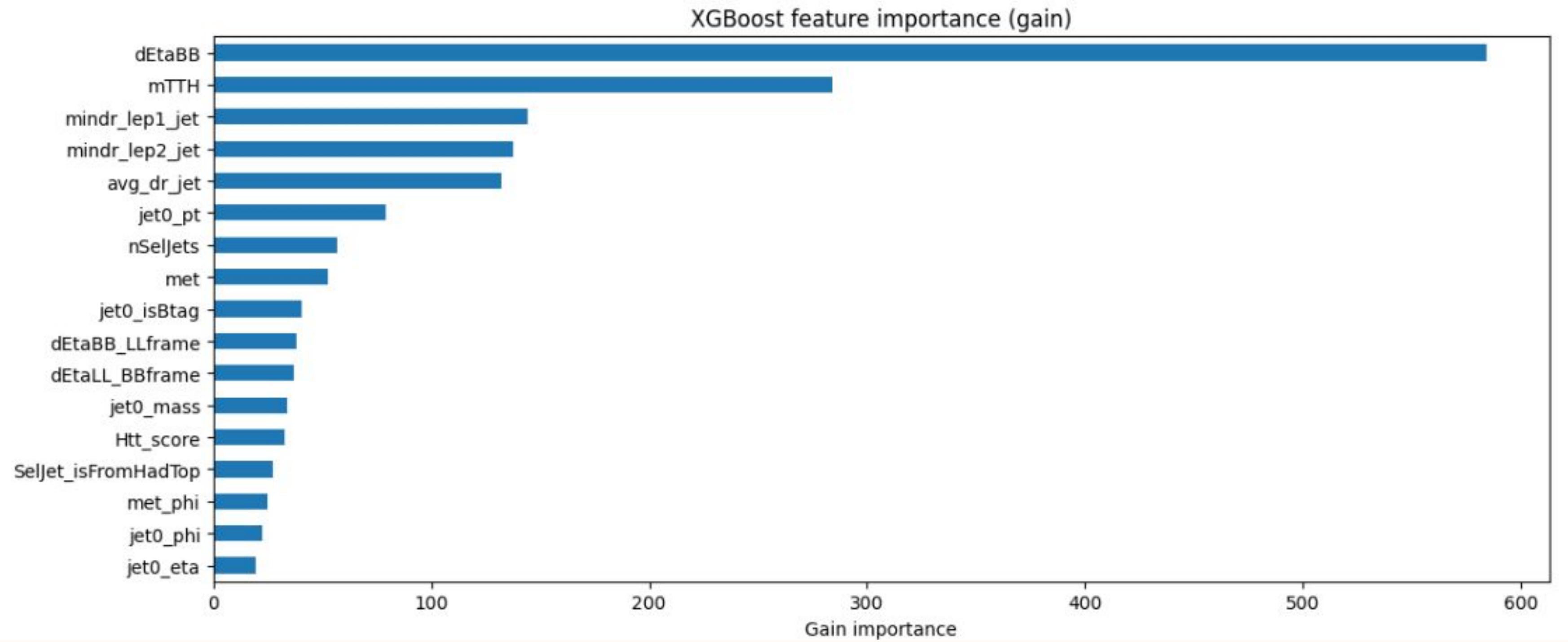
The Output: Loss function



Loss functions along the boosting rounds for the 2lss0tau (left) and 3l0tau (right) BDTs

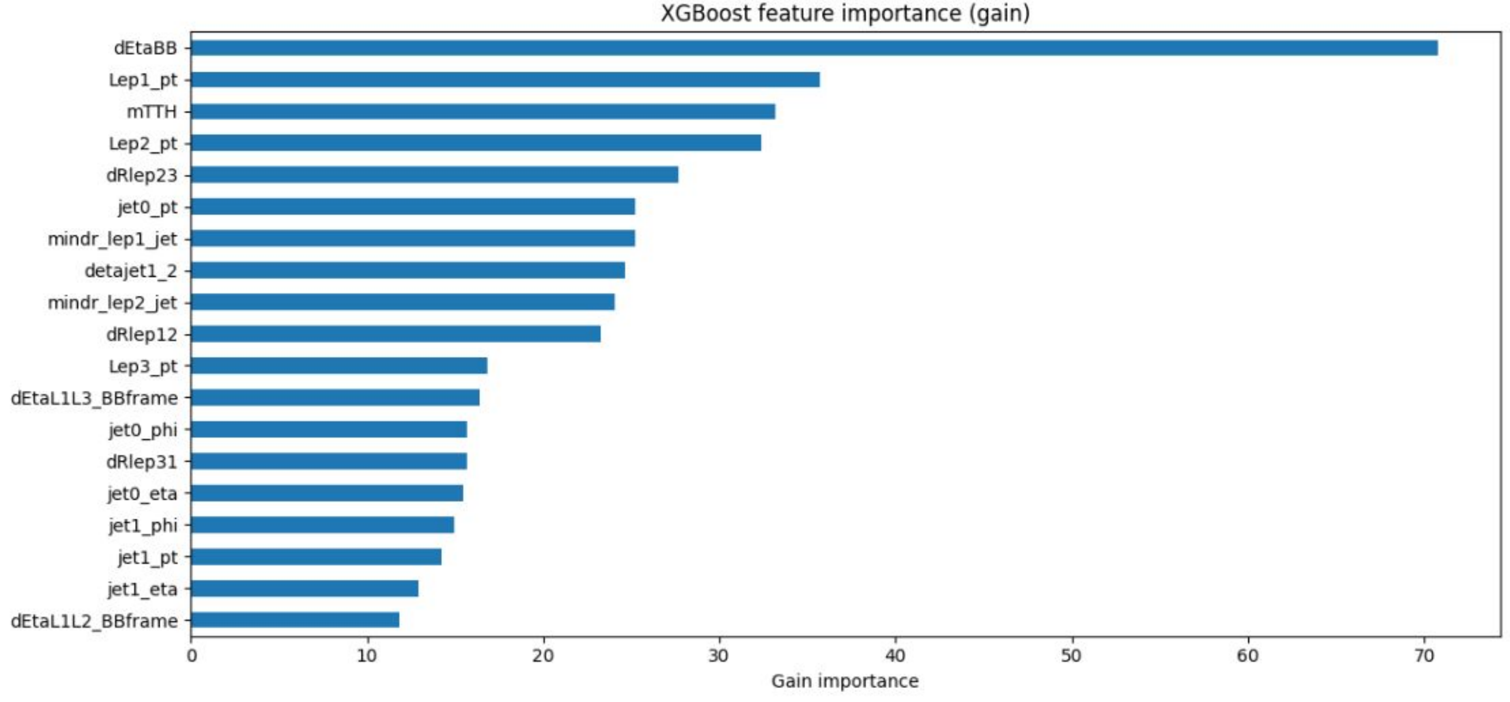
The Input Variables: Variable ranking in Run 3 (2lss0tau)

All features used for the 2lss0tau CP-BDT, with relative importance. The missing features are those relying on the Higgs-Jet tagger, the Htt_score, and the Seljet_btagDeepFlavB



The Input Variables: Variable ranking in Run 3 (3l0tau)

Features used for the 3l0tau BDT, with relative importance. All variables from Run2 were used, and extra variables regarding the properties of the jets were added



Variables from arXiv:2406.03950v2

observable	definition	frame
$p_{T,H}$	-	lab, $t\bar{t}$, $t\bar{t}H$
$\Delta\eta_{t\bar{t}}$	$\lvert\eta_t - \eta_{\bar{t}}\rvert$	lab, H, $t\bar{t}H$
$\Delta\phi_{t\bar{t}}$	$\lvert\phi_t - \phi_{\bar{t}}\rvert$	lab, H, $t\bar{t}H$
$m_{t\bar{t}}$	$(p_t + p_{\bar{t}})^2$	frame-invariant
$m_{t\bar{t}H}$	$(p_t + p_{\bar{t}} + p_H)^2$	frame-invariant
$\lvert\cos\theta^*\rvert$	$\frac{\lvert\boldsymbol{p}_t\cdot\boldsymbol{n}\rvert}{\lvert\boldsymbol{p}_t\rvert\cdot\lvert\boldsymbol{n}\rvert}$	$t\bar{t}$
b_1	$\frac{(\boldsymbol{p}_t\times\boldsymbol{n})\cdot(\boldsymbol{p}_{\bar{t}}\times\boldsymbol{n})}{p_{T,t}p_{T,\bar{t}}}$	all
b_2	$\frac{(\boldsymbol{p}_t\times\boldsymbol{n})\cdot(\boldsymbol{p}_{\bar{t}}\times\boldsymbol{n})}{\lvert\boldsymbol{p}_t\rvert\ \lvert\boldsymbol{p}_{\bar{t}}\rvert}$	all
b_3	$\frac{p_t^x\ p_{\bar{t}}^x}{p_{T,t}p_{T,\bar{t}}}$	all
b_4	$\frac{p_t^z\ p_{\bar{t}}^z}{\lvert\boldsymbol{p}_t\rvert\ \lvert\boldsymbol{p}_{\bar{t}}\rvert}$	all
ϕ_C	$\arccos\left(\frac{\lvert(\boldsymbol{p}_{p_1}\times\boldsymbol{p}_{p_2})\cdot(\boldsymbol{p}_t\times\boldsymbol{p}_{\bar{t}})\rvert}{\lvert\boldsymbol{p}_{p_1}\times\boldsymbol{p}_{p_2}\rvert\ \lvert\boldsymbol{p}_t\times\boldsymbol{p}_{\bar{t}}\rvert}\right)$	H

Except this one