National Research
Tomsk
State
University

Лаборатория
анализа данных
физики высоких энергий
Томского
государственного
университета

Measurement of differential cross-sections of a single top quark produced in association with a W boson with ATLAS at $\sqrt{s} = 13$ TeV

Progress Report

Neda Firoz

Goal: separate **tW (top+anti-top)** from **$t\bar{t}$** in the **1j1b dilepton** region

•**Inputs (9 variables used):**
bdt_centrality_ll_recalc_NOSYS,
bdt_delta_pT_ll_MET_recalc_NOSYS,
bdt_delta_pT_llb_MET_recalc_NOSYS,
bdt_eta_llMetB_recalc_NOSYS,
bdt_m_l1b_recalc_NOSYS,
bdt_m_l2b_recalc_NOSYS,
bdt_pT_llMetB_recalc_NOSYS,
bdt_pT_llb_recalc_NOSYS,
bdt_sum_ET_recalc_NOSYS.

•**Samples / tree:** all files' tree name is analysis

Signal: tW (top) + $\bar{t}$W (anti-top)
Background: $t\bar{t}$ (non-all-had)

•**Event weights:** auto-resolved to weight_mc_NOSYS * weight_pileup_NOSYS * globalTriggerEffSF_NOSYS.

•**Bad values:** any variable ≤ **−990** or non-finite is masked per event.

Training and Test Split

- 70% train / 30% test (by default test_size = 0.3).
- Stratified by class (stratify=y), so the signal:background ratio is preserved in both sets.
- Fixed random_state=42 → fully reproducible split.
- All input variables are standardized to zero mean and unit variance.

Classifiers Used

- GradientBoostingClassifier
- RandomForestClassifier
- AdaBoostClassifier
- KNeighborsClassifier (KNN).
- MLPClassifier (feed-forward neural network).
- GaussianNB (naive Bayes with Gaussian likelihoods).

TOOLKIT

- Python 3.12 on ATLAS LCG_108 stack
- ROOT I/O via uproot + pandas
- Machine learning: scikit-learn (GradientBoosting, RandomForest, AdaBoost, LogisticRegression, KNN, MLP, GaussianNB)
- Plotting: matplotlib in batch mode (ROC, distributions, PCA, correlations, feature importance)

Hyperparameter optimisation

- RandomizedSearchCV with 3-fold StratifiedKFold Objective: maximise ROC AUC
- Event weights were propagated as sample_weight when supported.
- For GaussianNB (no tunable parameters):Single global fit + 3-fold CV AUC for performance estimate

- We compare all models using weighted ROC AUC and pick the best one
- MLP ≈ 0.678 and GradientBoosting ≈ 0.682

Model Performance Comparison on Python Sklearn

Metric	Gradient Boosting (BDT)	MLP
Best CV AUC	0.674793	0.671154
Test AUC	0.670624	**0.678083**
Threshold	0.5000	0.5000
Accuracy	**0.991551**	0.988750
True Positives (TP)	0.000	1023.228
False Positives (FP)	0.000	14639.035
True Negatives (TN)	4,819,801.582	4,805,162.547
False Negatives (FN)	41,068.097	40,044.869

Article’s Report on BDT

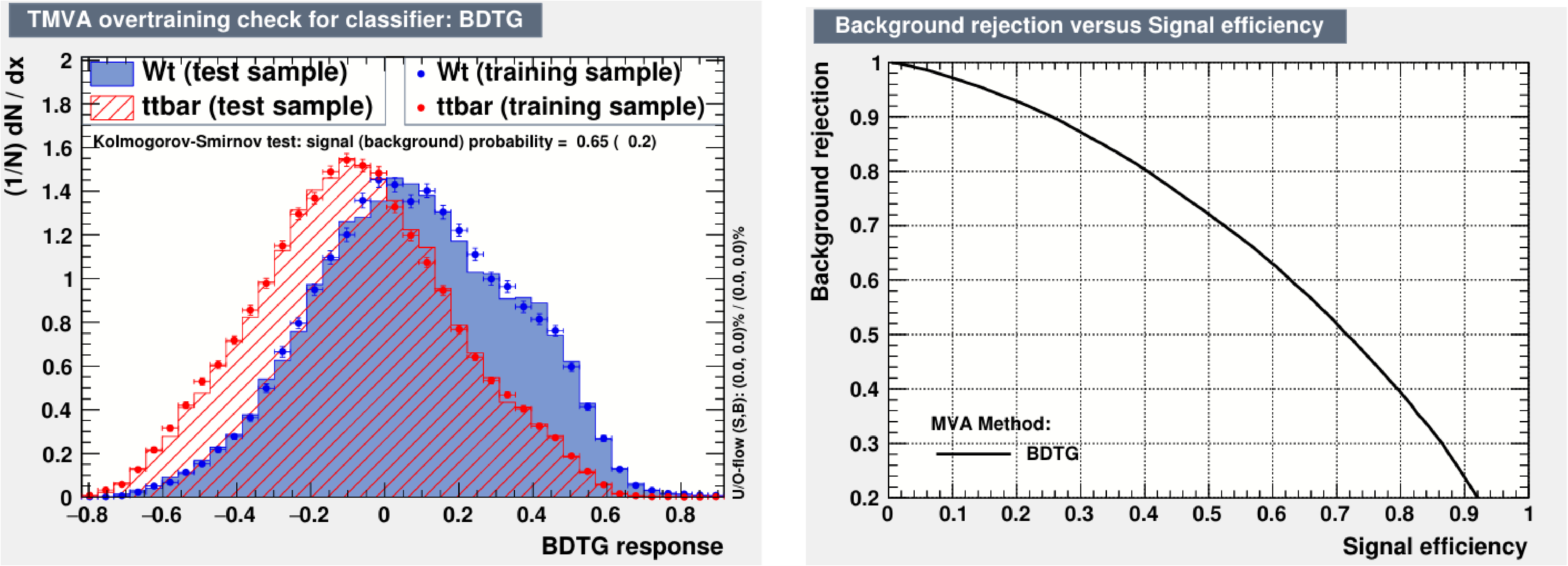
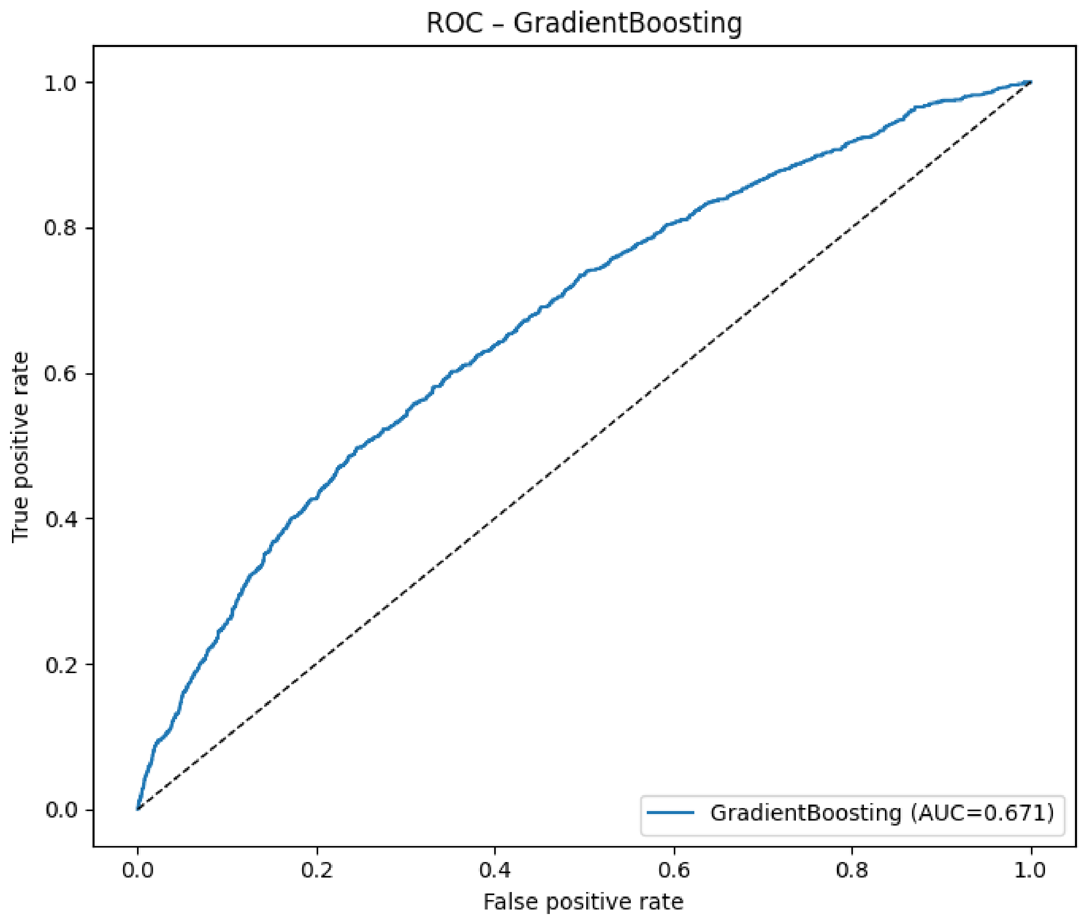
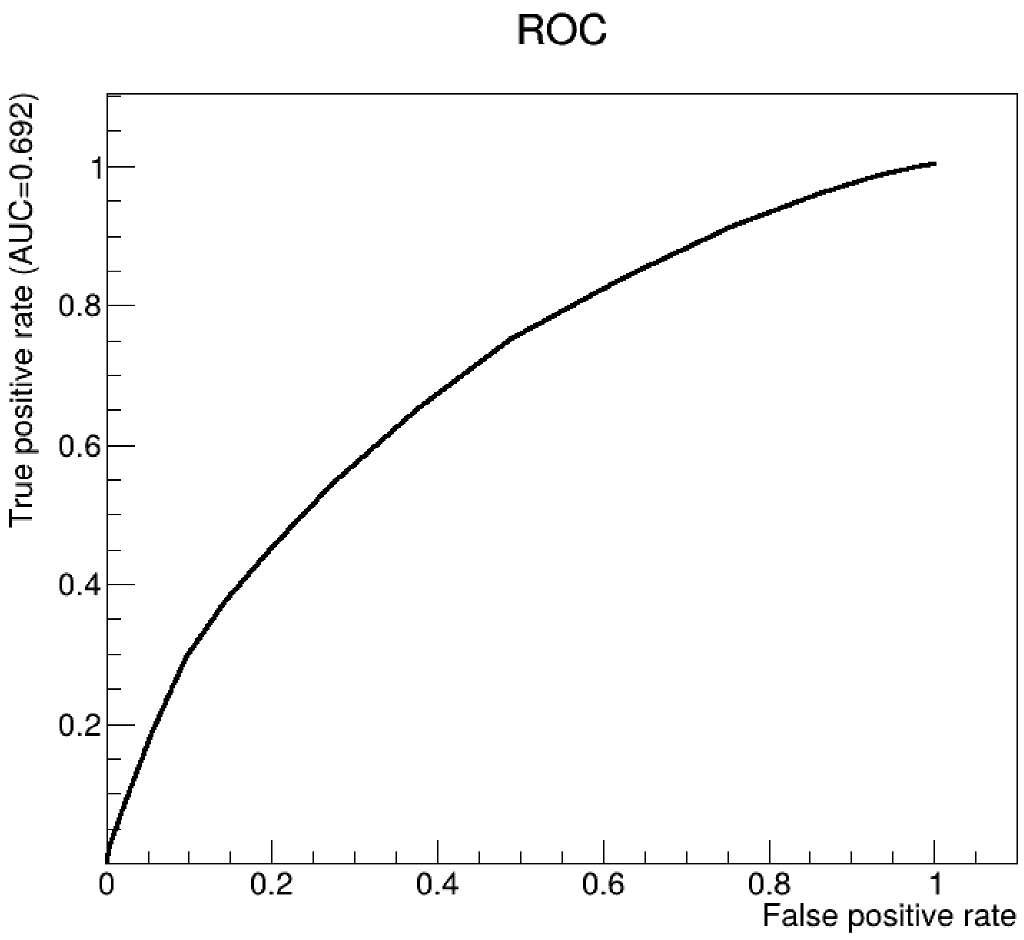


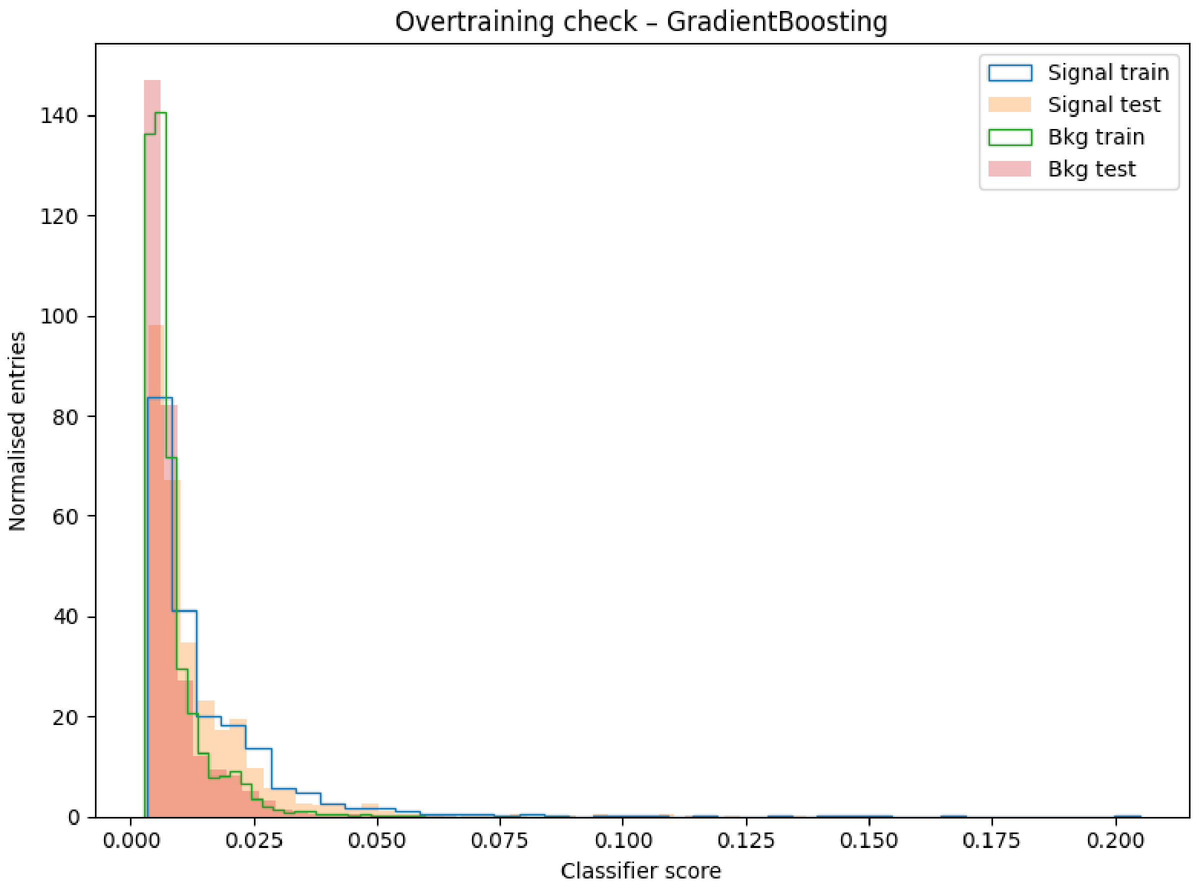
Figure 9: Comparison of test/training sample distributions and background rejection factor versus signal efficiency.

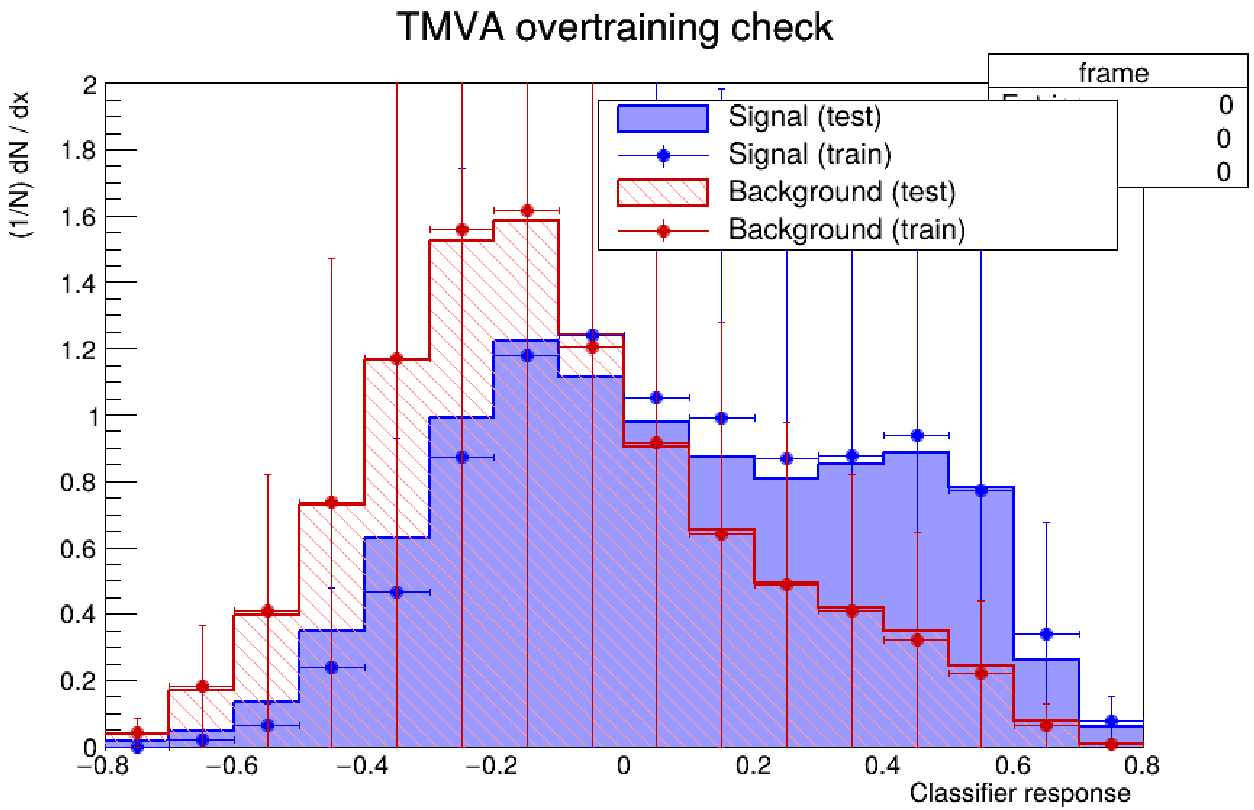
ROC of BDT from sklearn and TMVA



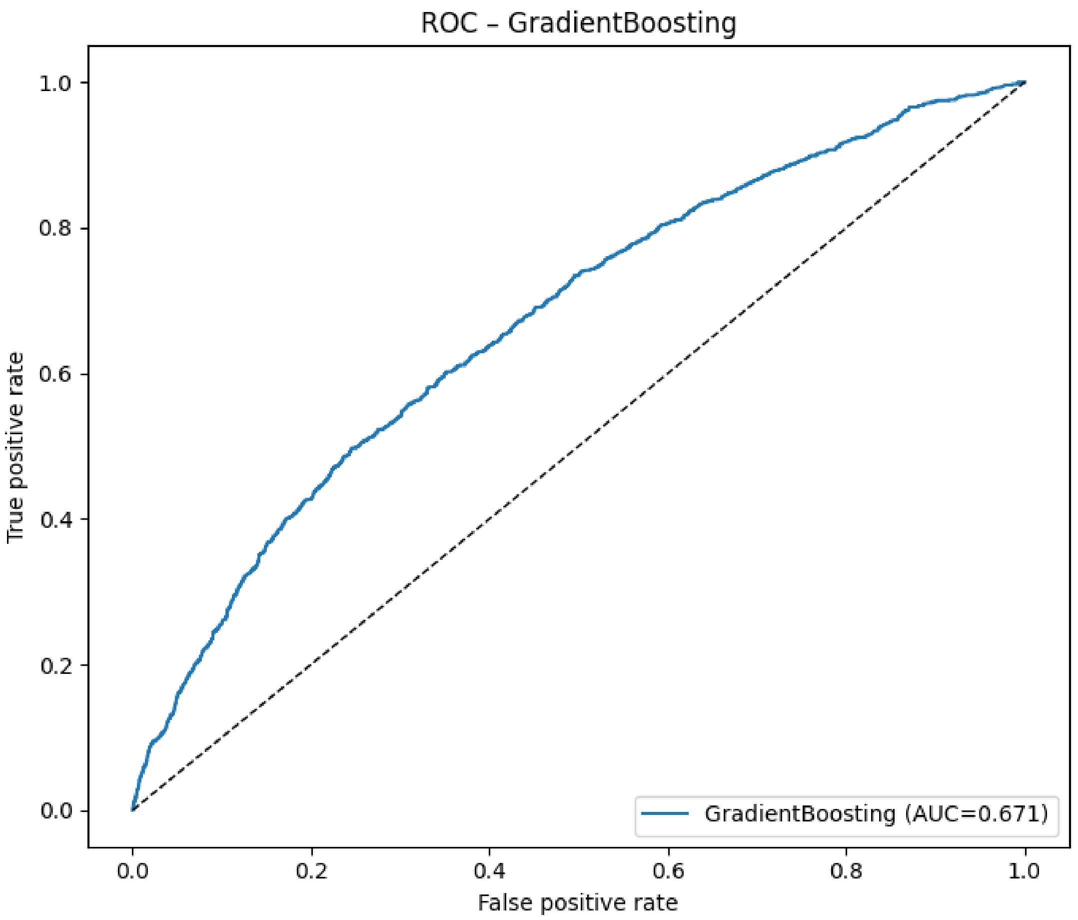


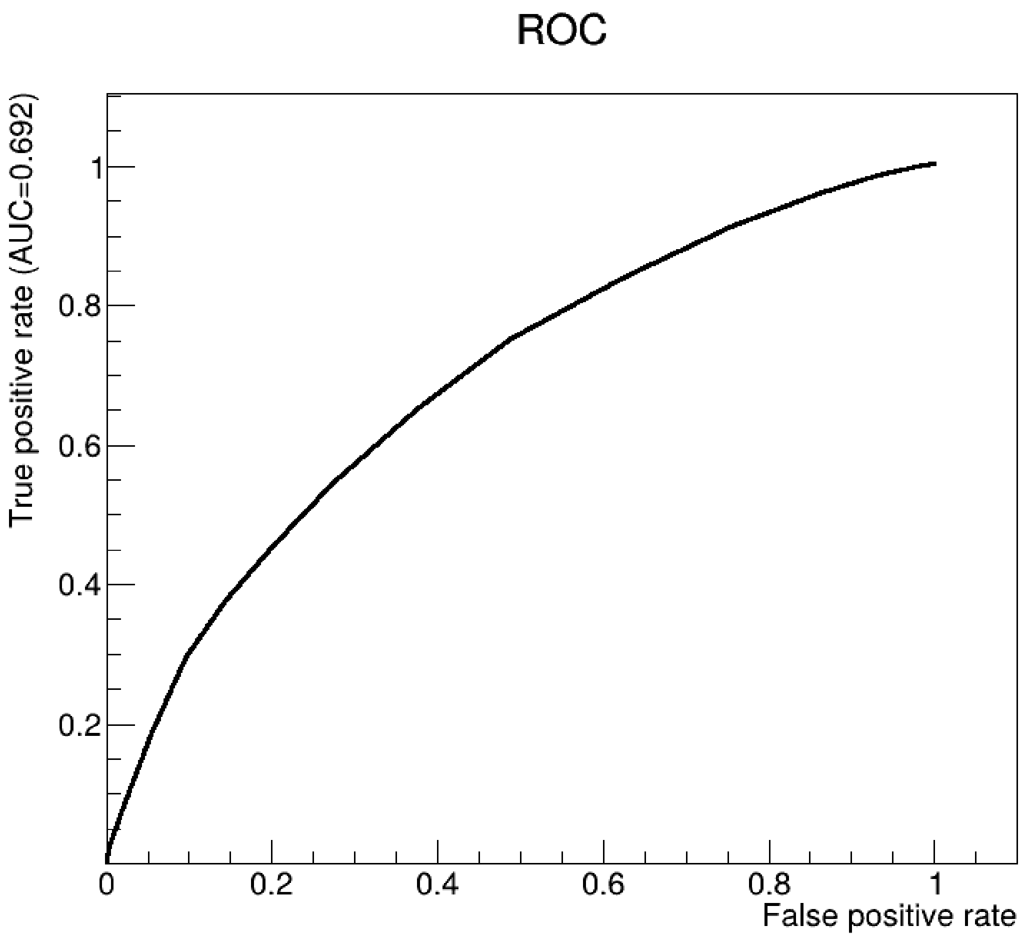
Gradient Boosting Algorithm from R and BDT from TMVA



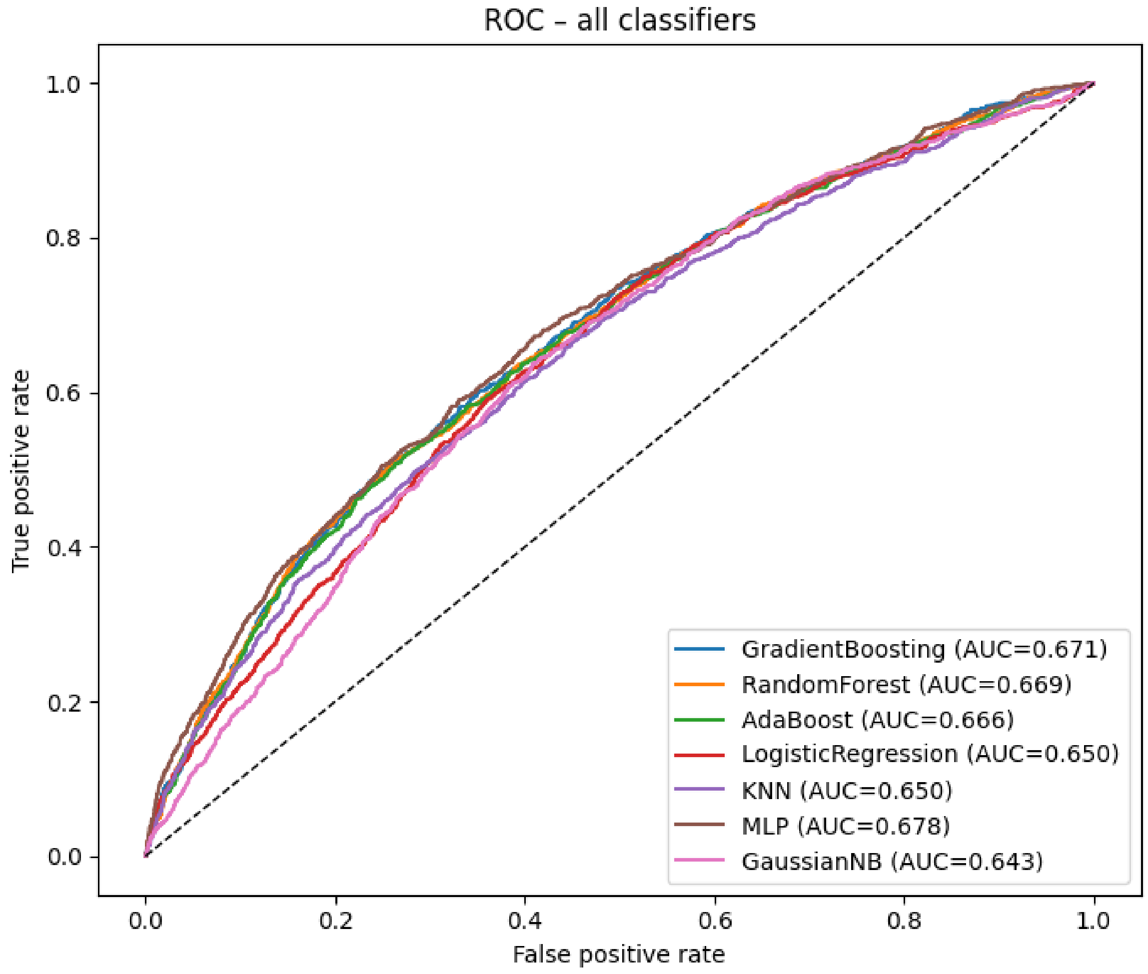


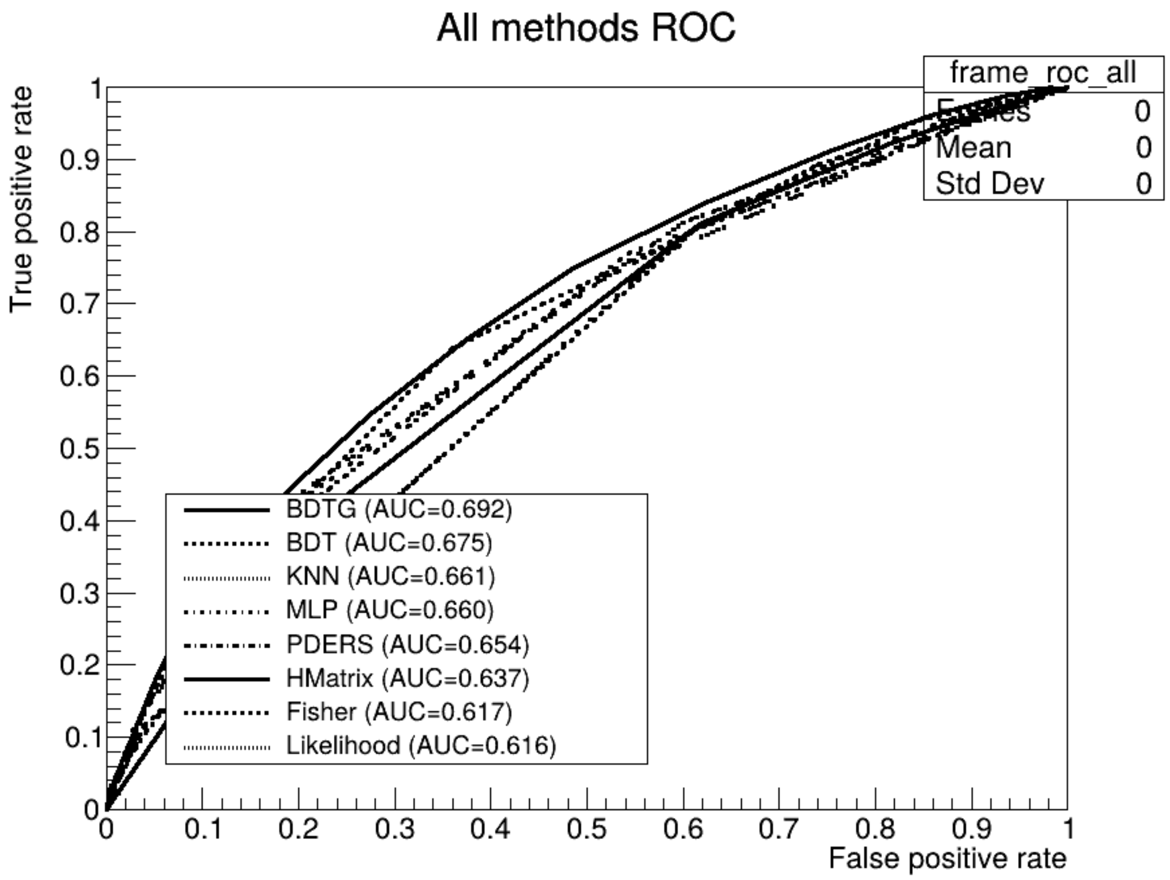
ROC of BDT from R and TMVA



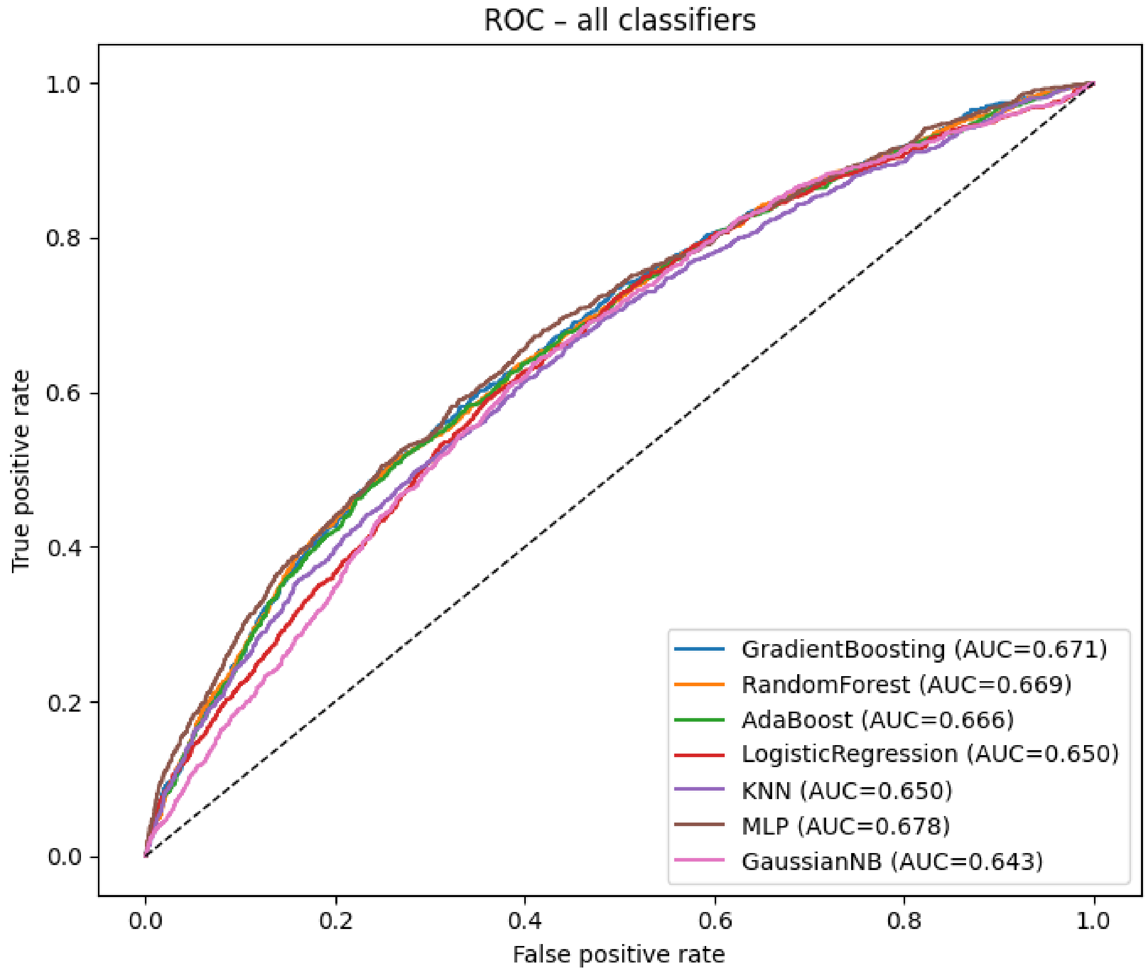


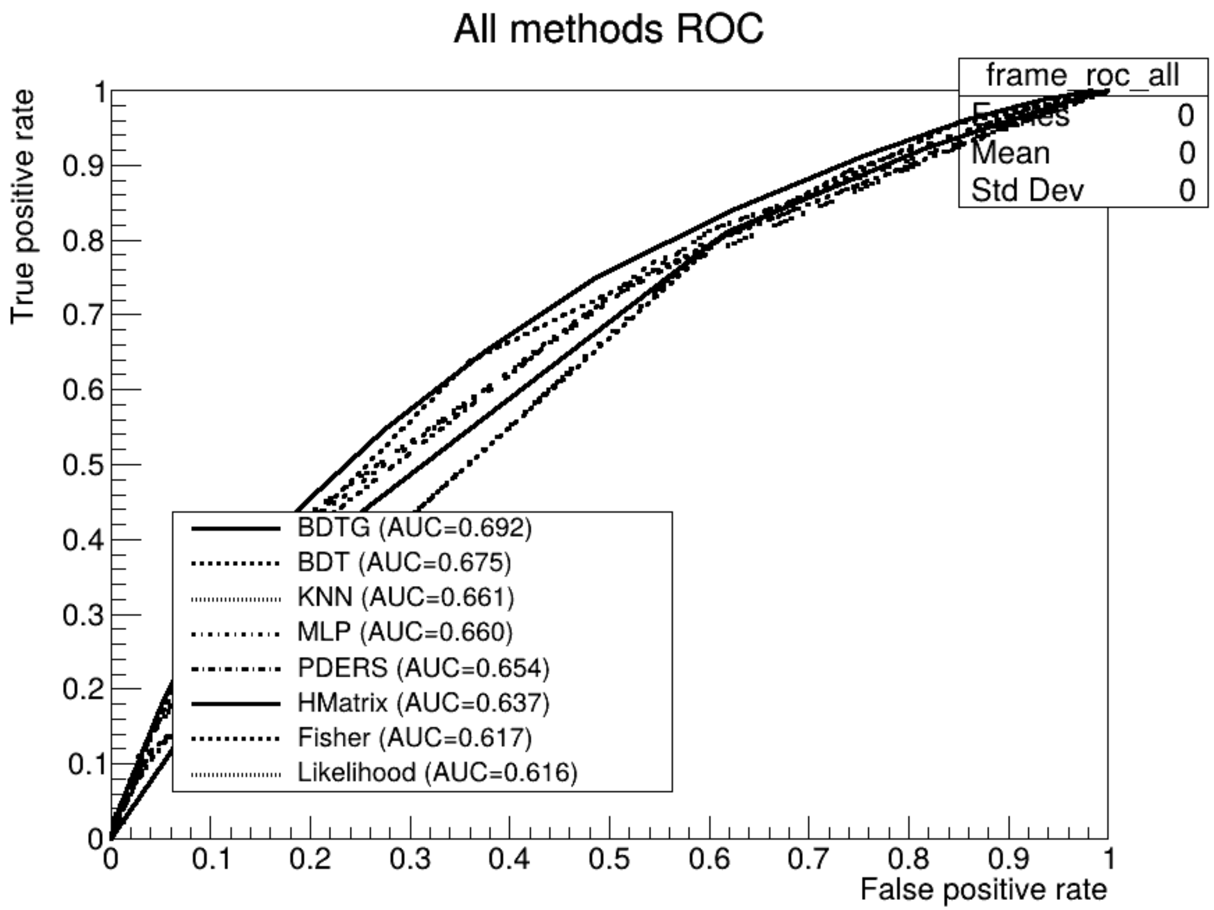
All classifiers ROC from sklearn and TMVA



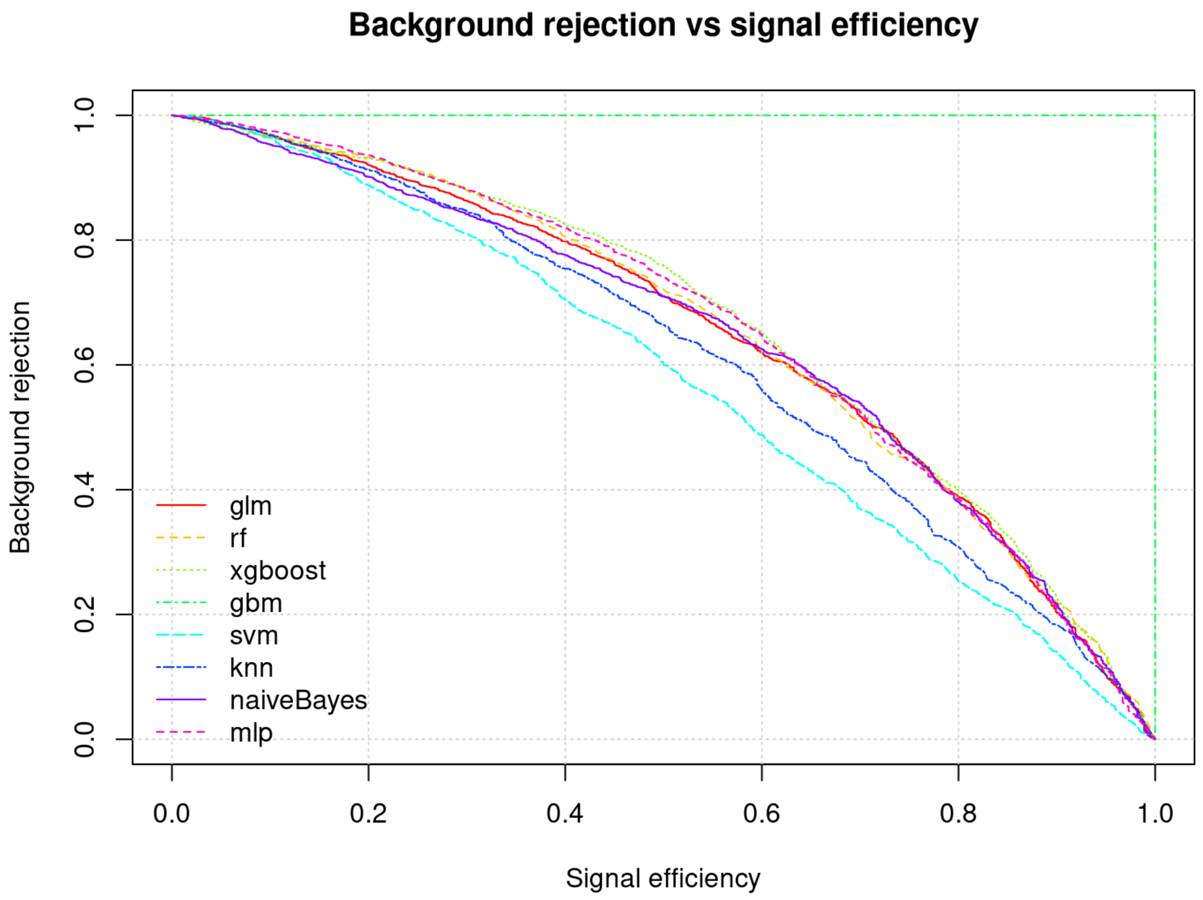


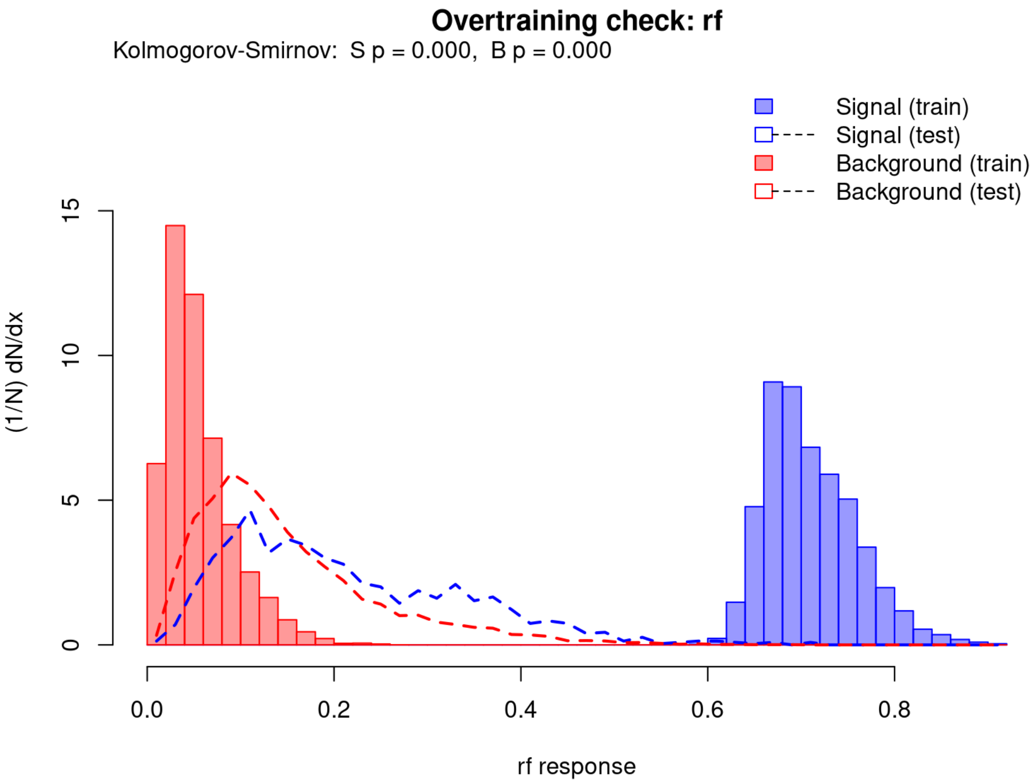
All classifiers ROC compared from R and TMVA environments

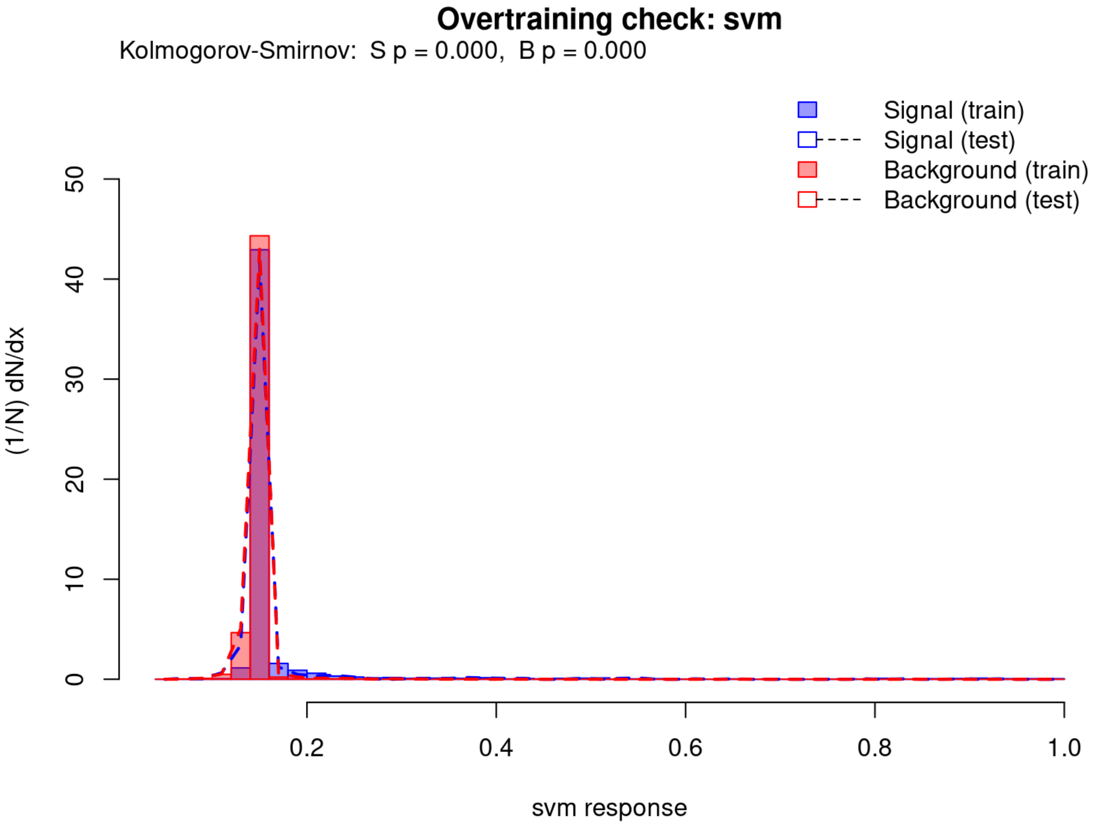


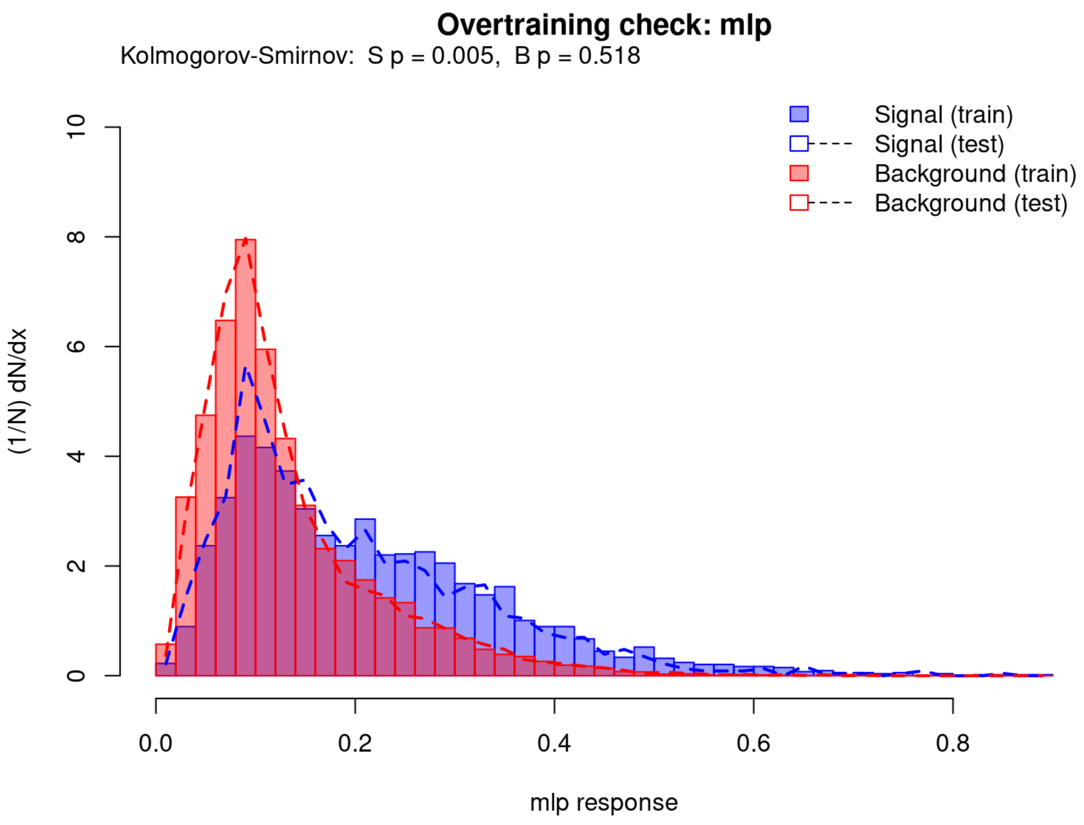


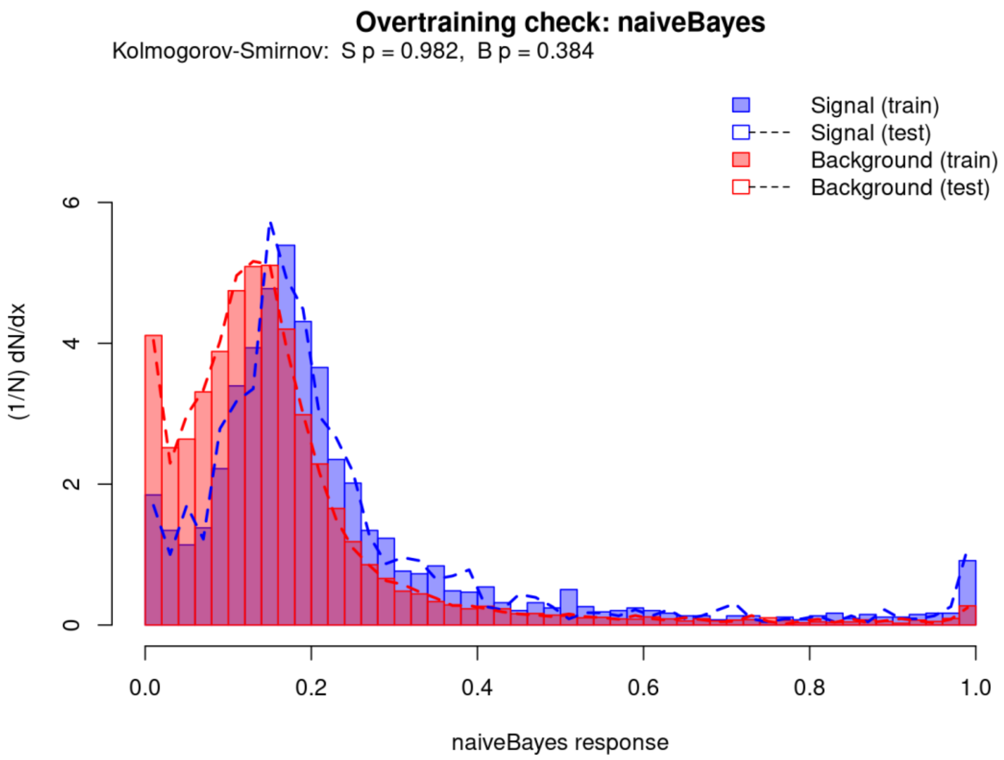
R based results

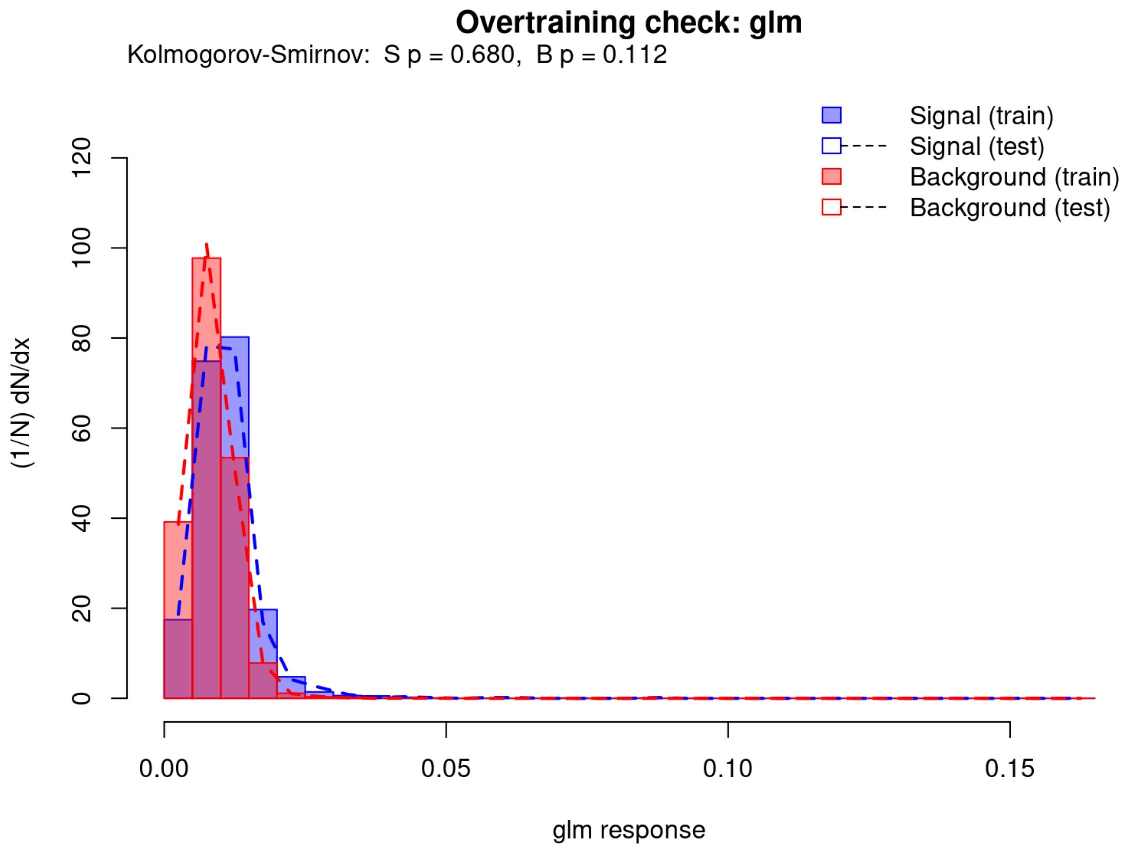



Overtraining check: rf
Kolmogorov-Smirnov: S p = 0.000, B p = 0.000
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
5
10
15
0.0
0.2
0.4
0.6
0.8
rf response

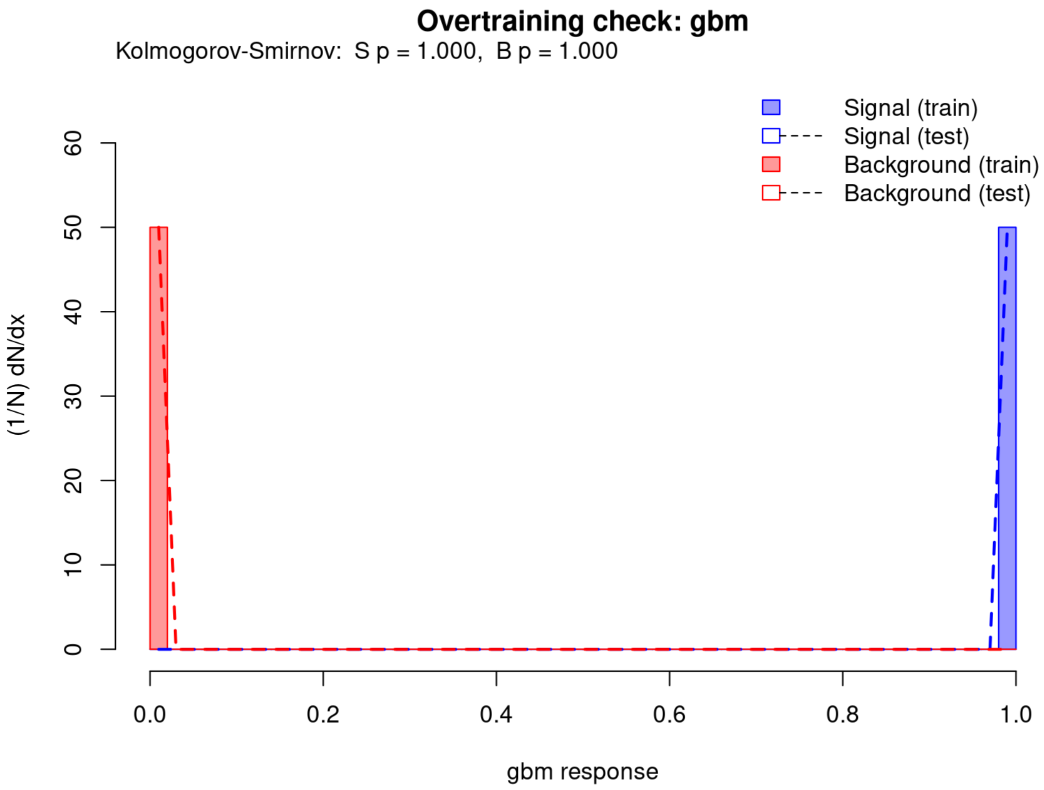

Overtraining check: svm
Kolmogorov-Smirnov: S p = 0.000, B p = 0.000
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
10
20
30
40
50
0.2
0.4
0.6
0.8
1.0
svm response


Overtraining check: mlp
Kolmogorov-Smirnov: S p = 0.005, B p = 0.518
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
mlp response

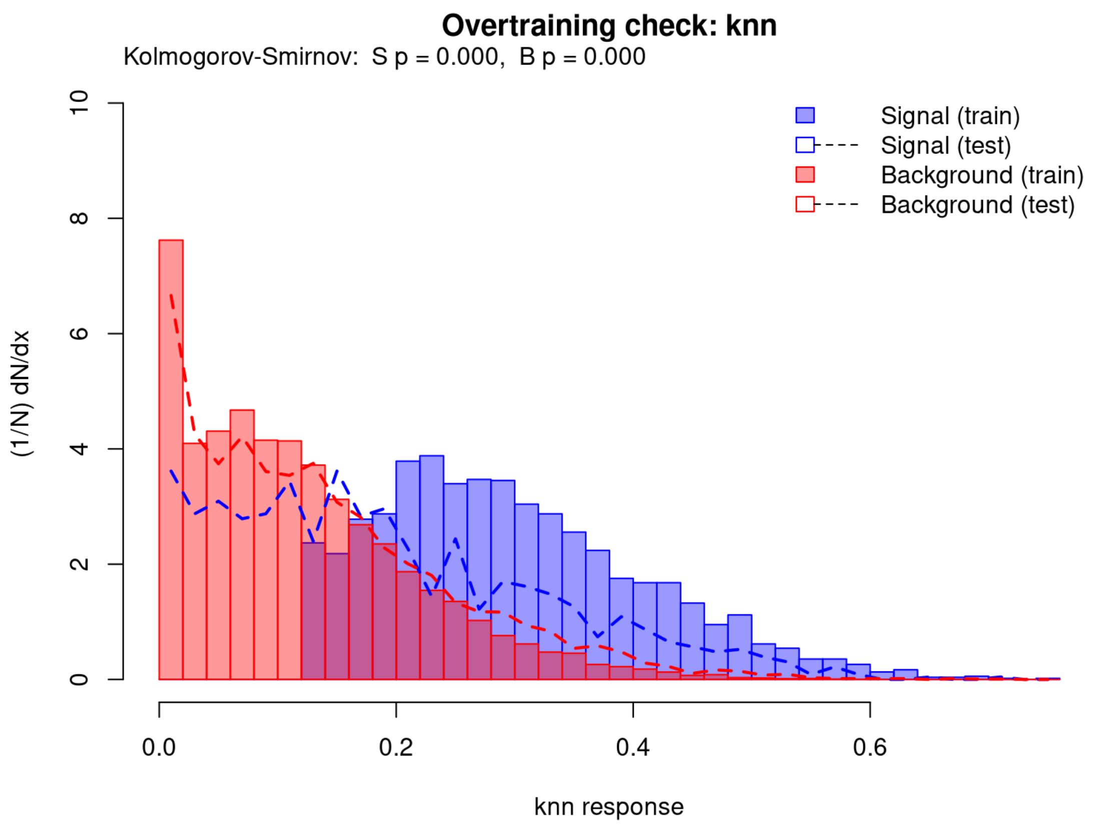

Overtraining check: naiveBayes
Kolmogorov-Smirnov: S p = 0.982, B p = 0.384
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
naiveBayes response

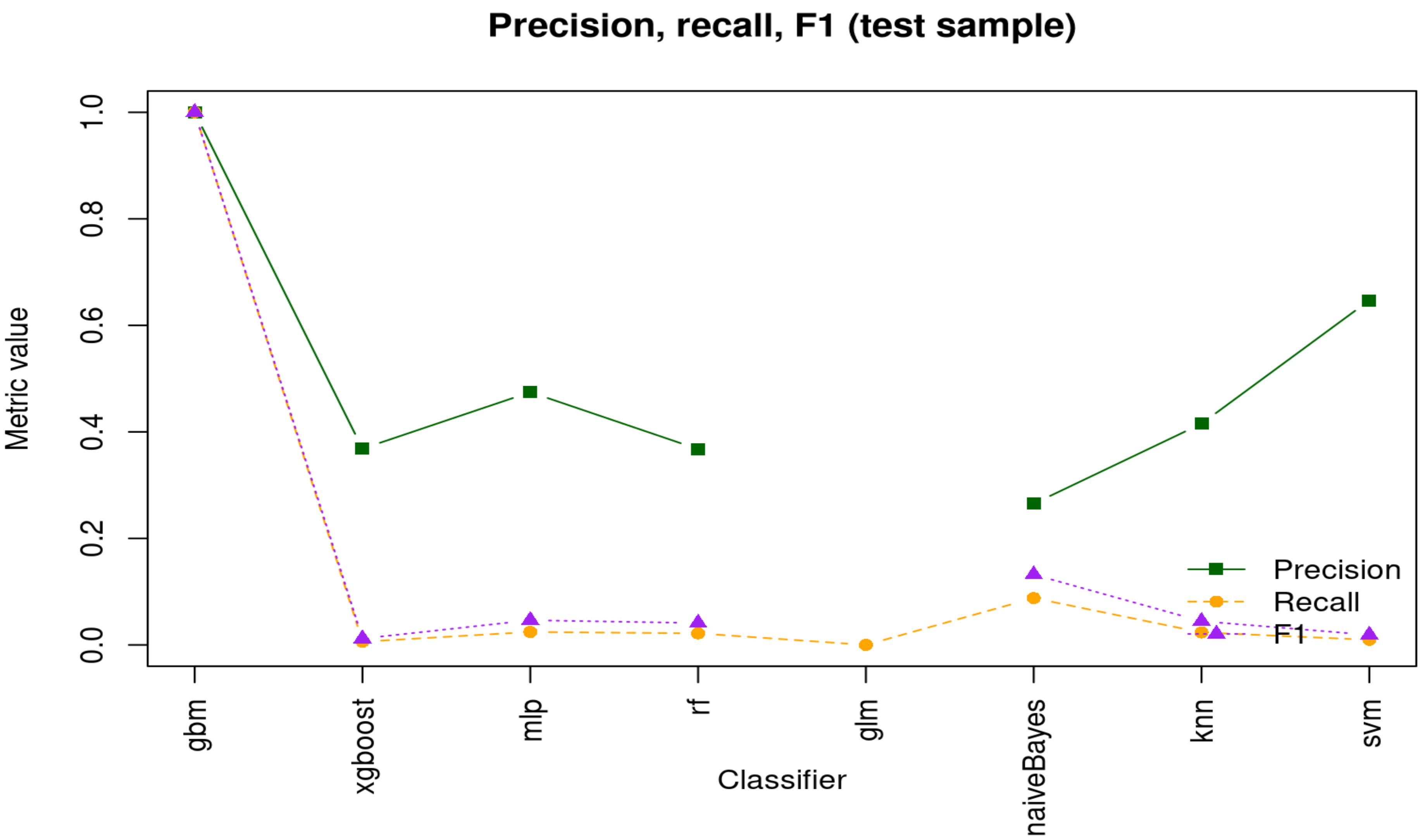


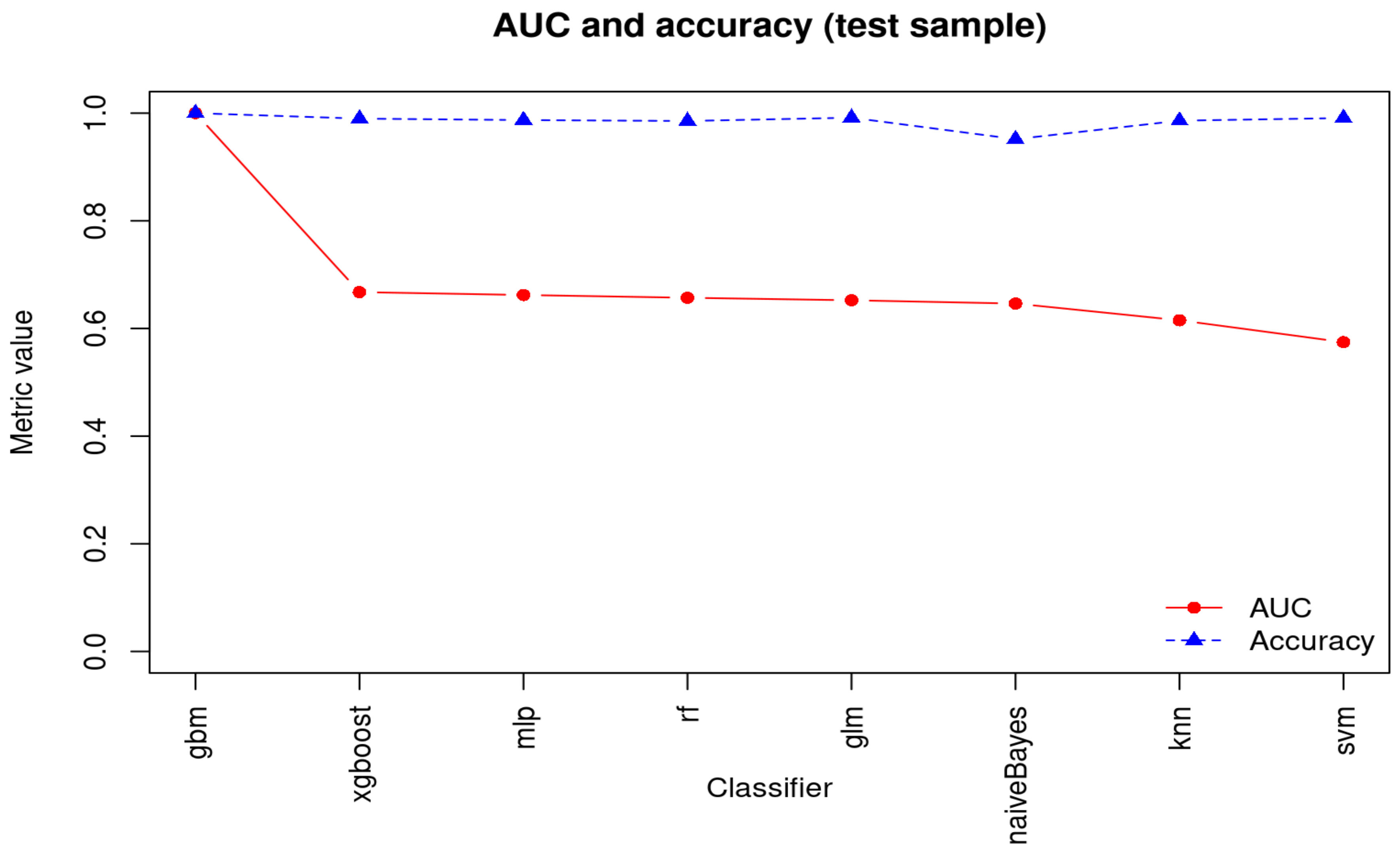
Overtraining check: glm
Kolmogorov-Smirnov: S p = 0.680, B p = 0.112
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
20
40
60
80
100
120
0.00
0.05
0.10
0.15
glm response



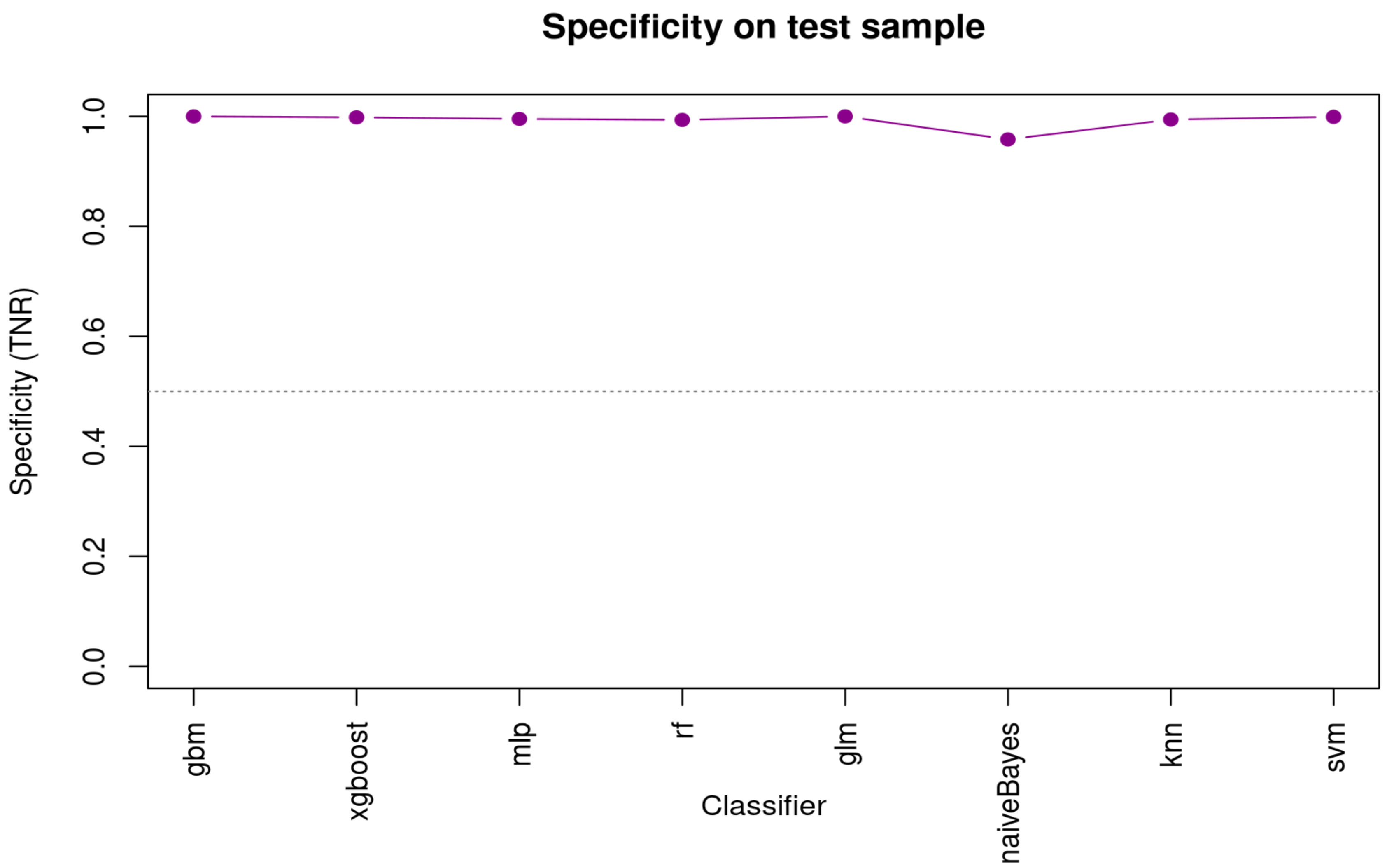
Overtraining check: gbm
Kolmogorov-Smirnov: S p = 1.000, B p = 1.000
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
gbm response

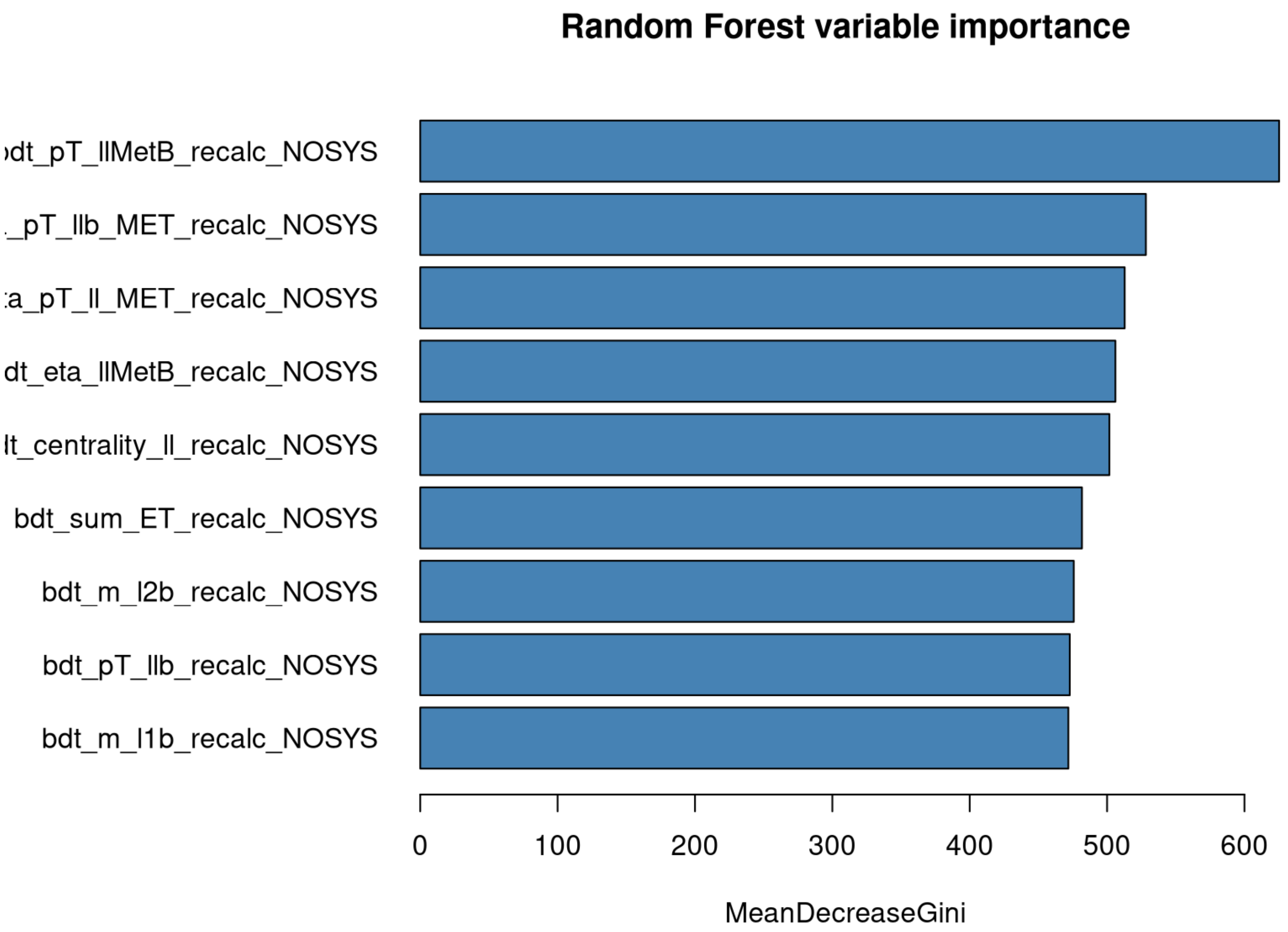

Overtraining check: knn
Kolmogorov-Smirnov: S p = 0.000, B p = 0.000
Signal (train)
Signal (test)
Background (train)
Background (test)
(1/N) dN/dx
0
2
4
6
8
10
0.0
0.2
0.4
0.6
knn response

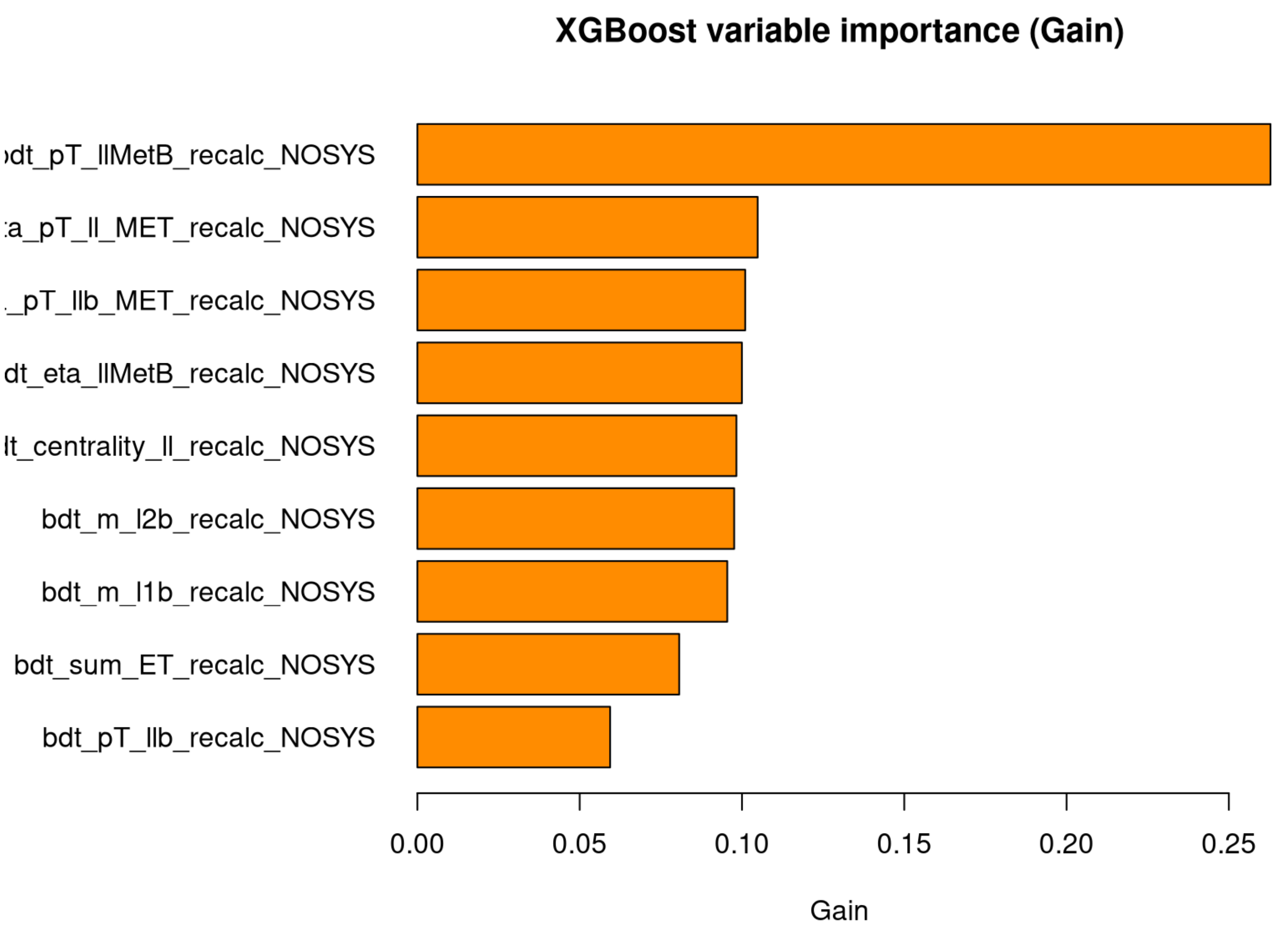

Precision, recall, F1 (test sample)
Metric value
0.0
0.2
0.4
0.6
0.8
1.0
gbm
xgboost
mlp
rf
glm
naiveBayes
knn
svm
Classifier
Precision
Recall
F1

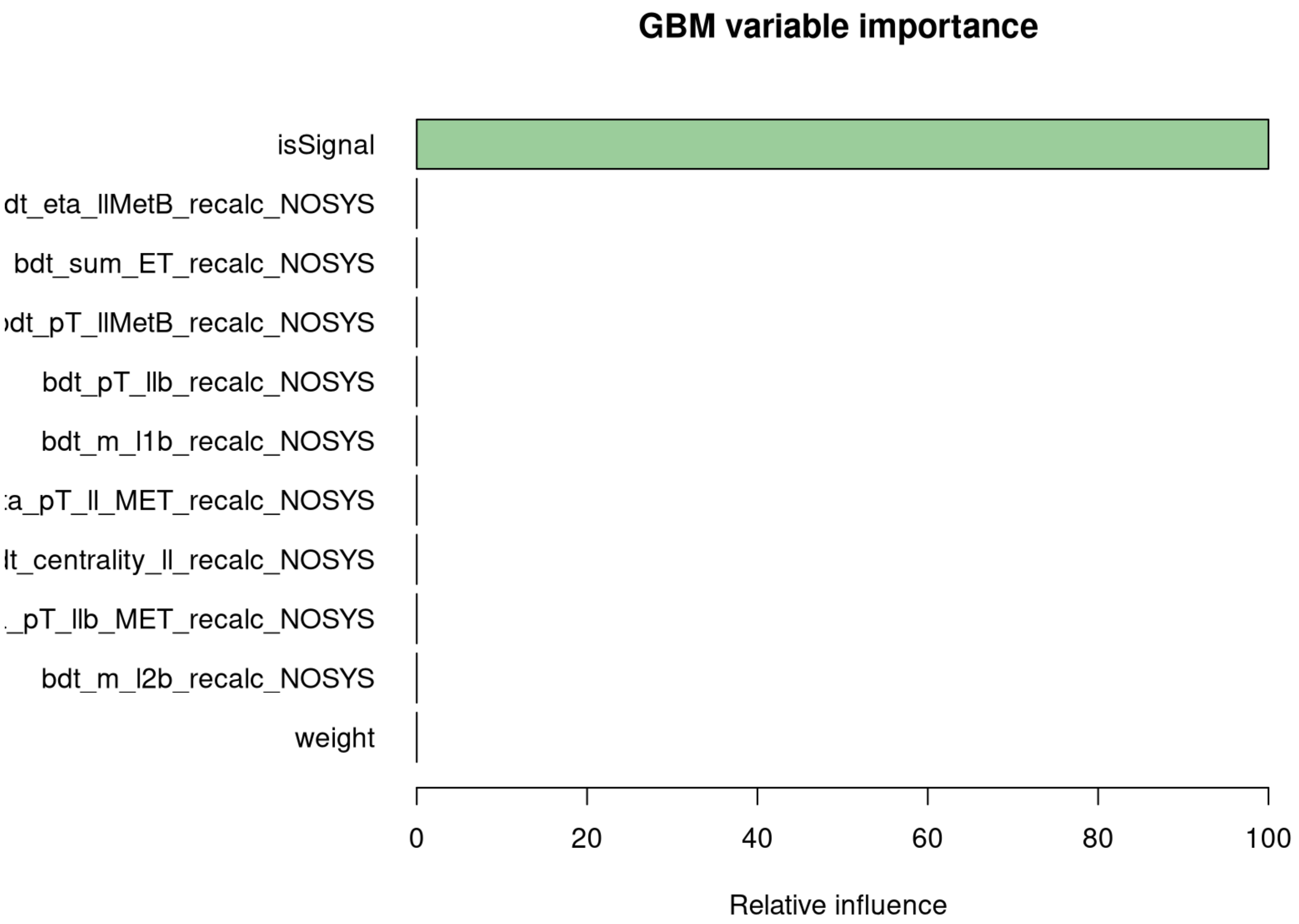


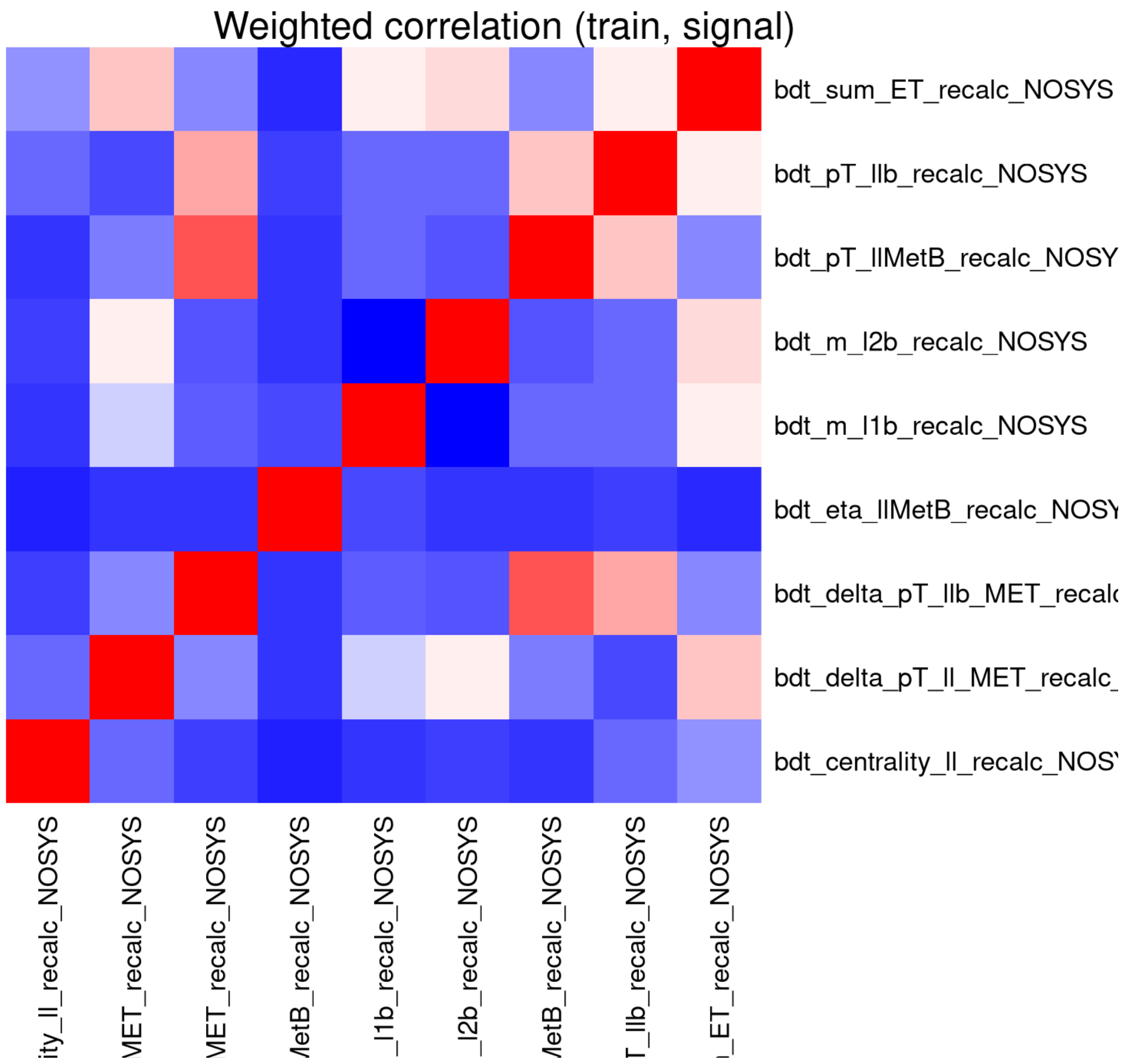
AUC and accuracy (test sample)
Metric value
0.0
0.2
0.4
0.6
0.8
1.0
gbm
xgboost
mlp
rf
glm
naiveBayes
knn
svm
Classifier
AUC
Accuracy


Specificity on test sample
Specificity (TNR)
0.0
0.2
0.4
0.6
0.8
1.0
gbm
xgboost
mlp
rf
glm
naiveBayes
knn
svm
Classifier

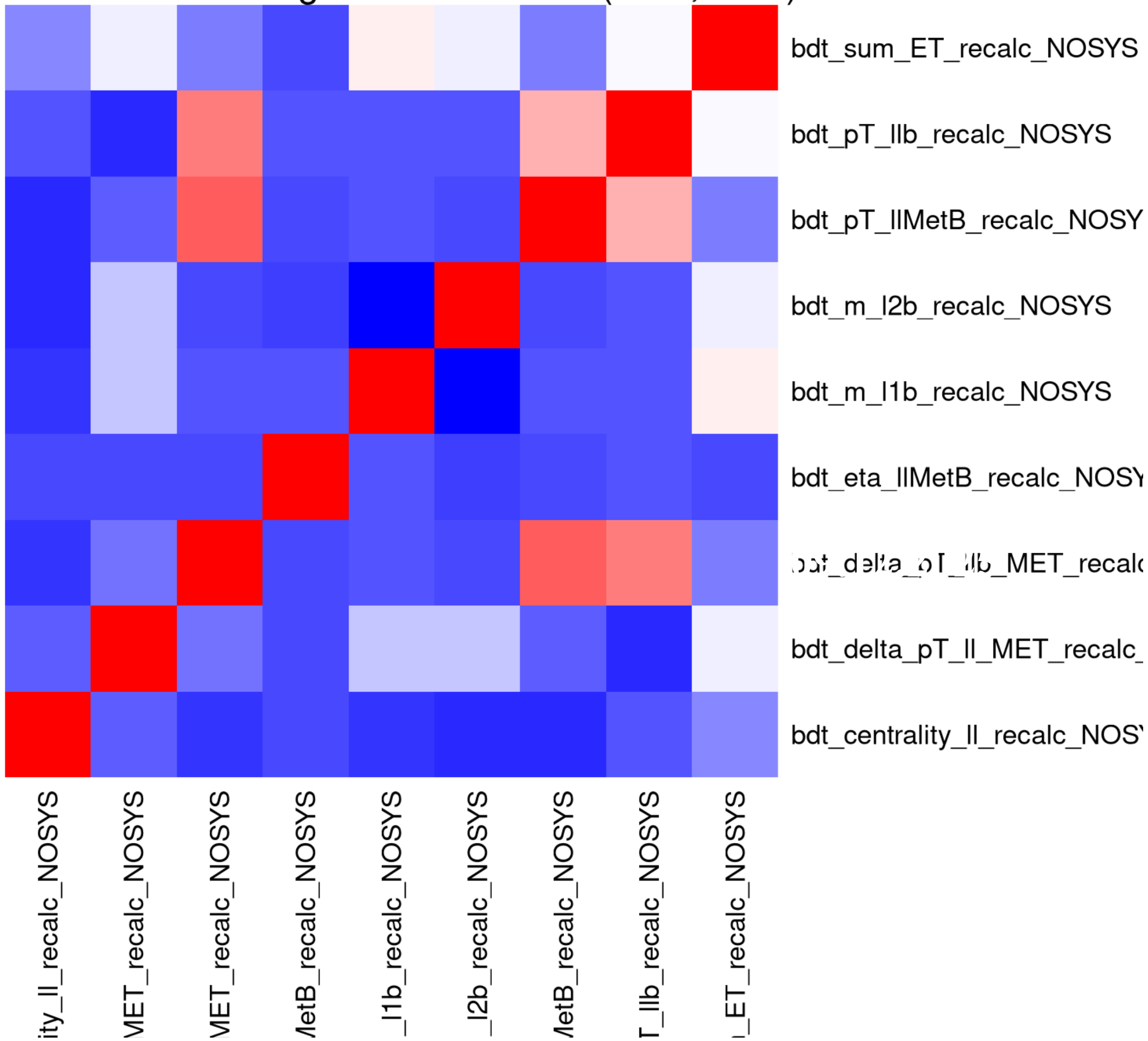

Random Forest variable importance
›dt_pT_llMetB_recalc_NOSYS
._pT_llb_MET_recalc_NOSYS
:a_pT_ll_MET_recalc_NOSYS
dt_eta_llMetB_recalc_NOSYS
lt_centrality_ll_recalc_NOSYS
bdt_sum_ET_recalc_NOSYS
bdt_m_l2b_recalc_NOSYS
bdt_pT_llb_recalc_NOSYS
bdt_m_l1b_recalc_NOSYS
0
100
200
300
400
500
600
MeanDecreaseGini

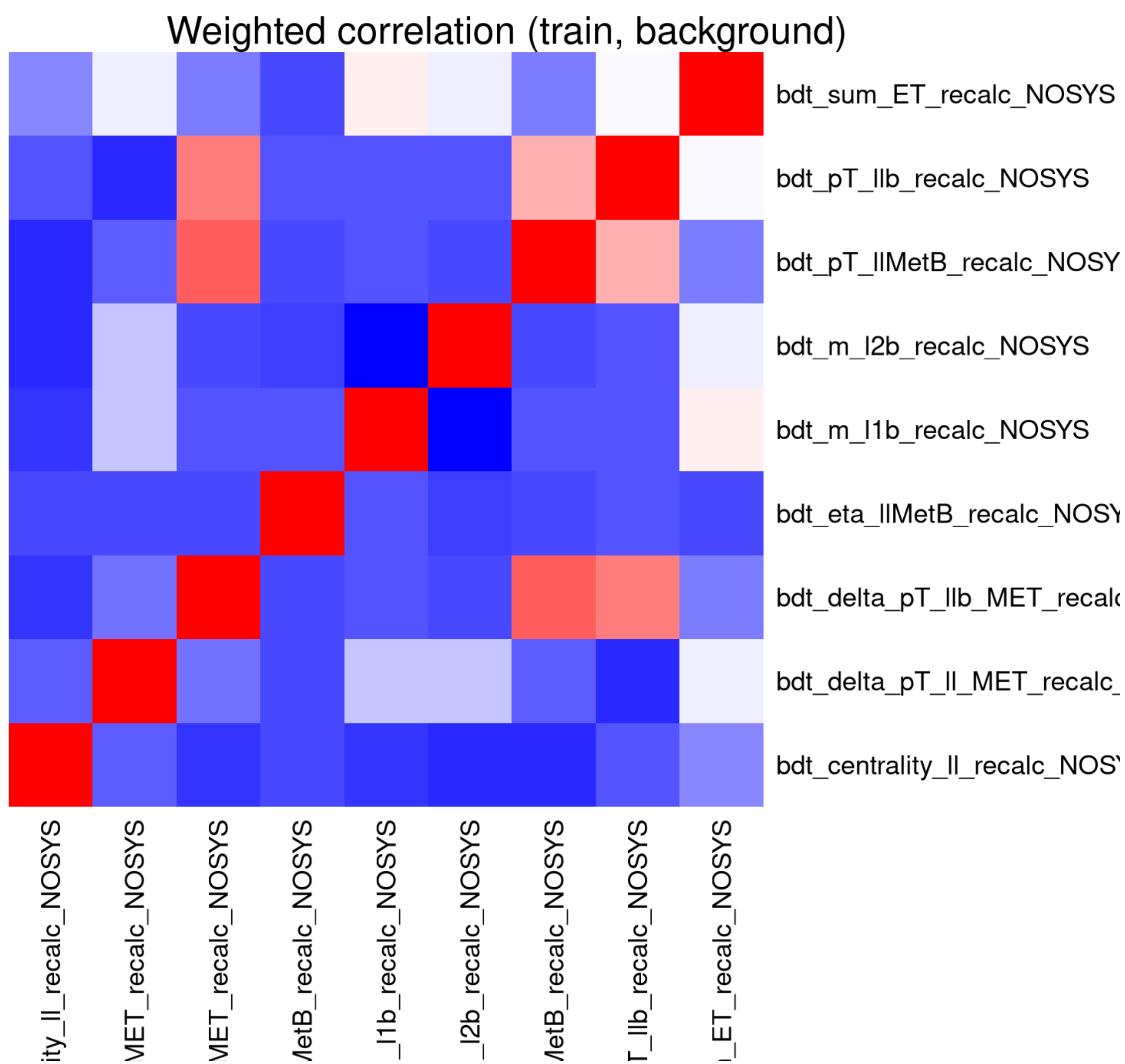

XGBoost variable importance (Gain)
›dt_pT_llMetB_recalc_NOSYS
:a_pT_ll_MET_recalc_NOSYS
._pT_llb_MET_recalc_NOSYS
dt_eta_llMetB_recalc_NOSYS
lt_centrality_ll_recalc_NOSYS
bdt_m_l2b_recalc_NOSYS
bdt_m_l1b_recalc_NOSYS
bdt_sum_ET_recalc_NOSYS
bdt_pT_llb_recalc_NOSYS
0.00
0.05
0.10
0.15
0.20
0.25
Gain


GBM variable importance
isSignal
dt_eta_llMetB_recalc_NOSYS
bdt_sum_ET_recalc_NOSYS
odt_pT_llMetB_recalc_NOSYS
bdt_pT_llb_recalc_NOSYS
bdt_m_l1b_recalc_NOSYS
:a_pT_ll_MET_recalc_NOSYS
lt_centrality_ll_recalc_NOSYS
._pT_llb_MET_recalc_NOSYS
bdt_m_l2b_recalc_NOSYS
weight
0
20
40
60
80
100
Relative influence

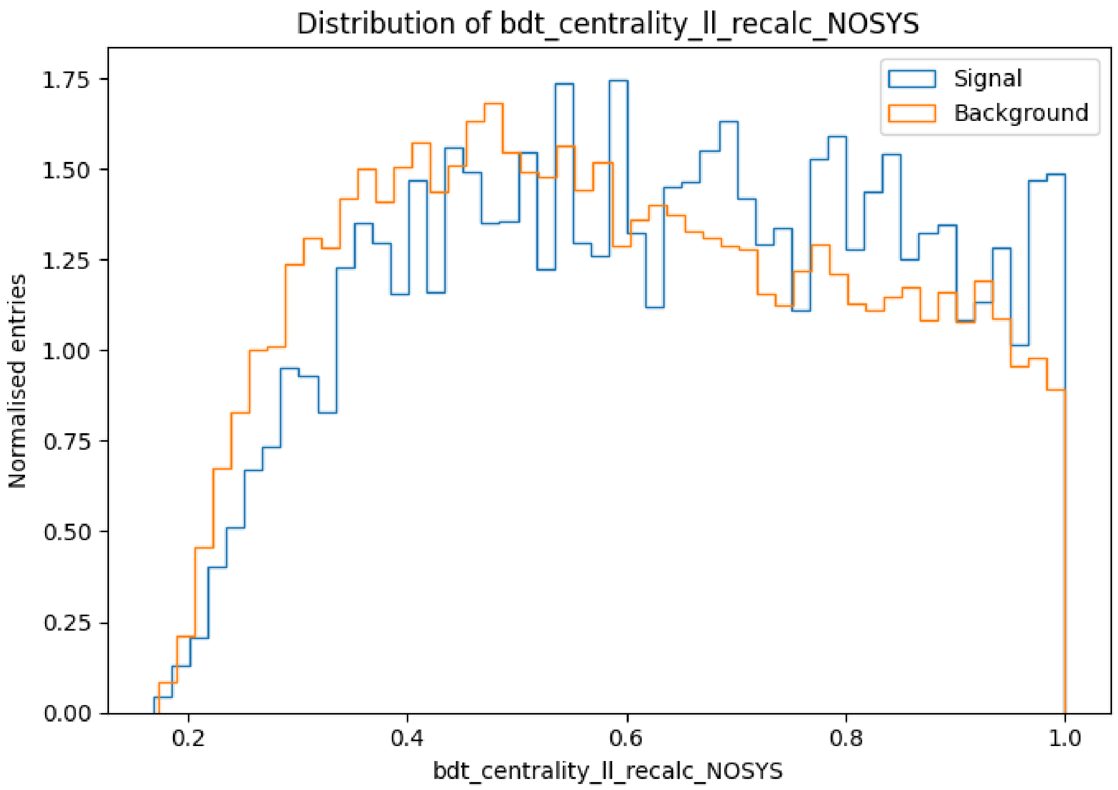

Weighted correlation (train, signal)
bdt_sum_ET_recalc_NOSYS
bdt_pT_llb_recalc_NOSYS
bdt_pT_llMetB_recalc_NOSY
bdt_m_l2b_recalc_NOSYS
bdt_m_l1b_recalc_NOSYS
bdt_eta_llMetB_recalc_NOSY
bdt_delta_pT_llb_MET_recal
bdt_delta_pT_ll_MET_recalc_
bdt_centrality_ll_recalc_NOS
ity_ll_recalc_NOSYS
MET_recalc_NOSYS
MET_recalc_NOSYS
letB_recalc_NOSYS
_l1b_recalc_NOSYS
_l2b_recalc_NOSYS
letB_recalc_NOSYS
_llb_recalc_NOSYS
_ET_recalc_NOSYS

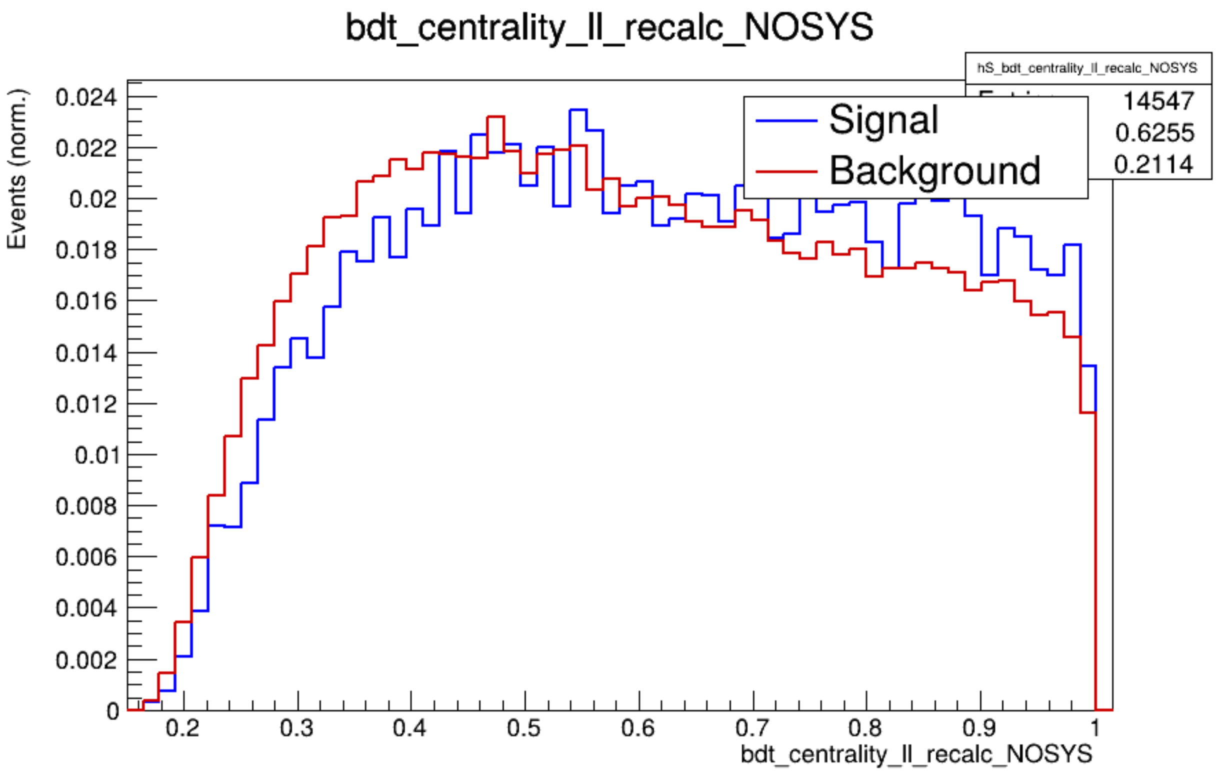
Weighted correlation (train, S+B)

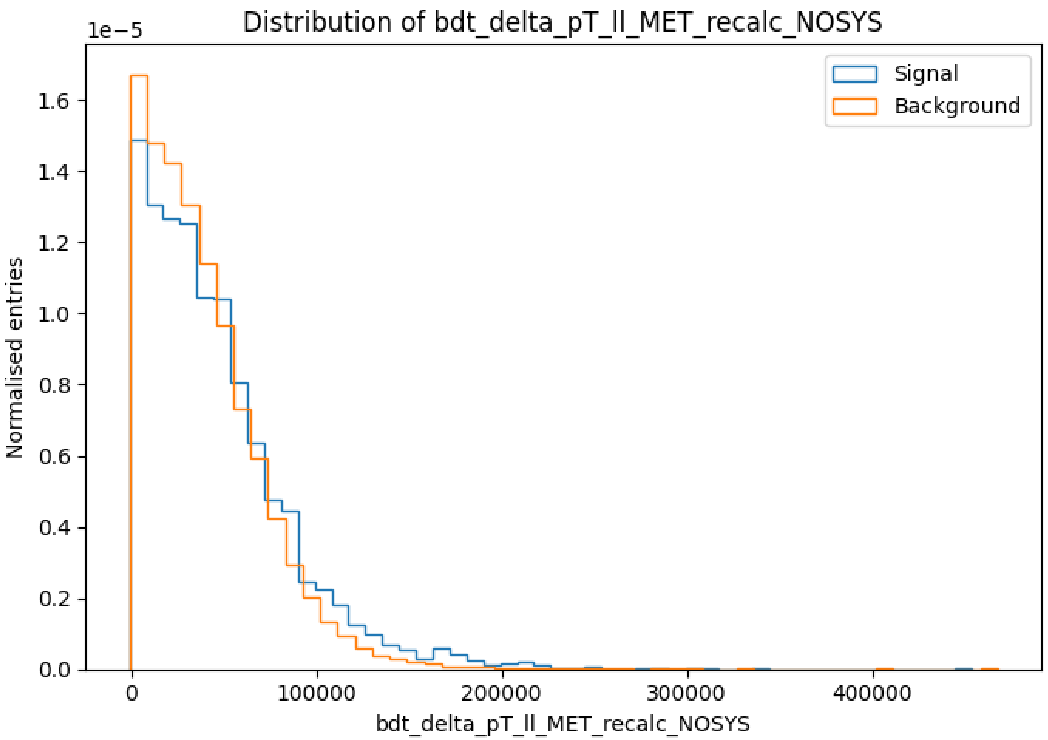



Weighted correlation (train, background)
bdt_sum_ET_recalc_NOSYS
bdt_pT_llb_recalc_NOSYS
bdt_pT_llMetB_recalc_NOSY
bdt_m_l2b_recalc_NOSYS
bdt_m_l1b_recalc_NOSYS
bdt_eta_llMetB_recalc_NOSY
bdt_delta_pT_llb_MET_recalc
bdt_delta_pT_ll_MET_recalc_
bdt_centrality_ll_recalc_NOSY
ity_ll_recalc_NOSYS
MET_recalc_NOSYS
MET_recalc_NOSYS
letB_recalc_NOSYS
_l1b_recalc_NOSYS
_l2b_recalc_NOSYS
letB_recalc_NOSYS
T_llb_recalc_NOSYS
_ET_recalc_NOSYS

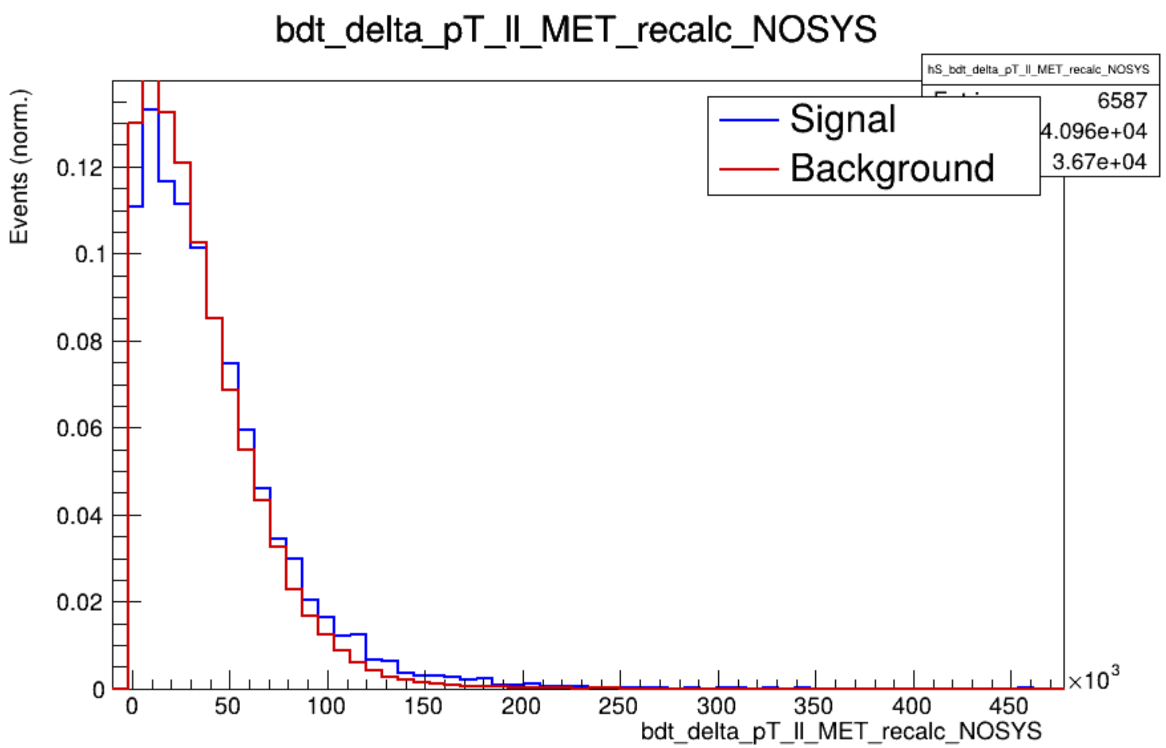
Variable separations comparison of plots in sklearn and TMVA environment


Distribution of bdt_centrality_ll_recalc_NOSYS
Signal
Background
Normalised entries
bdt_centrality_ll_recalc_NOSYS
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0.2
0.4
0.6
0.8
1.0

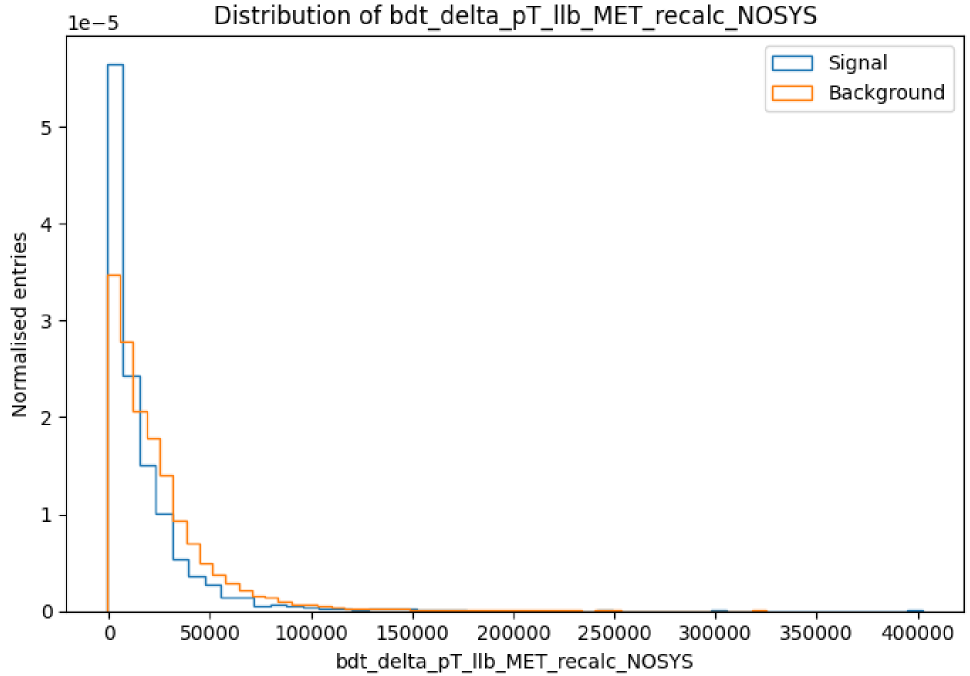

bdt_centrality_ll_recalc_NOSYS
hS_bdt_centrality_ll_recalc_NOSYS
14547
0.6255
0.2114
Signal
Background
Events (norm.)
bdt_centrality_ll_recalc_NOSYS
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
0.02
0.022
0.024
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

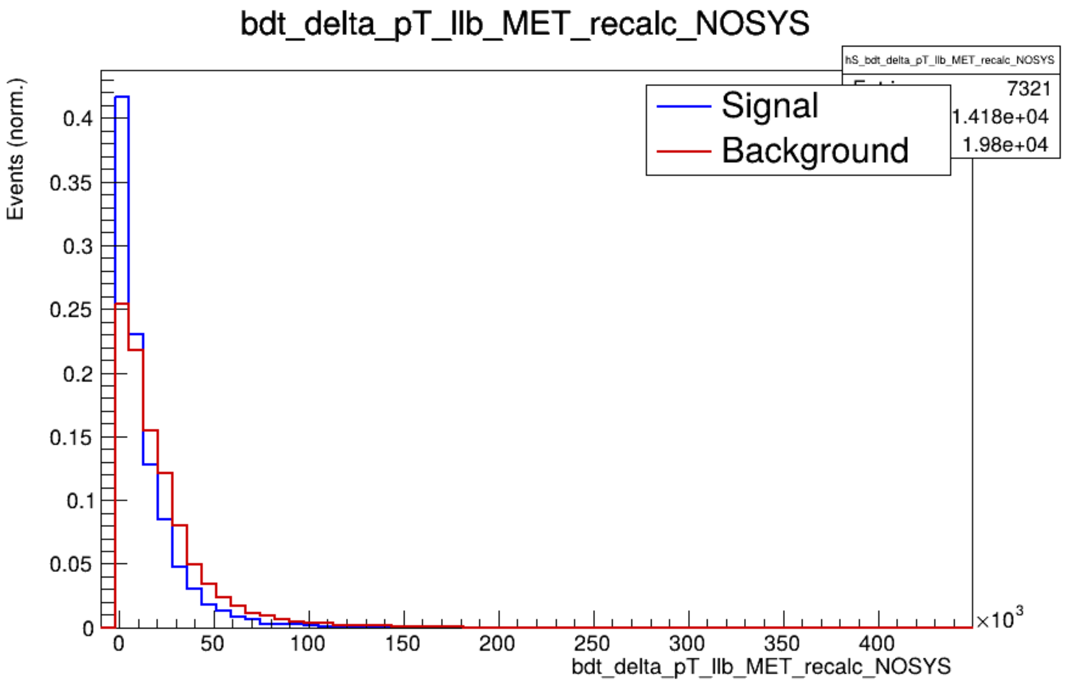


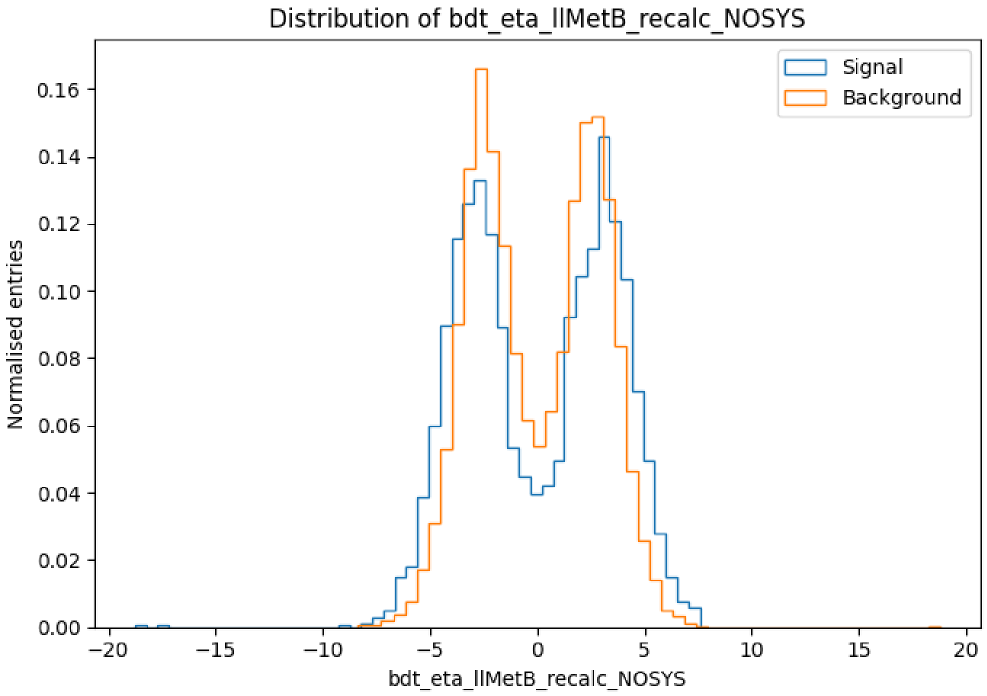
Distribution of bdt_delta_pT_ll_MET_recalc_NOSYS
1e−5
Signal
Background
Normalised entries
bdt_delta_pT_ll_MET_recalc_NOSYS

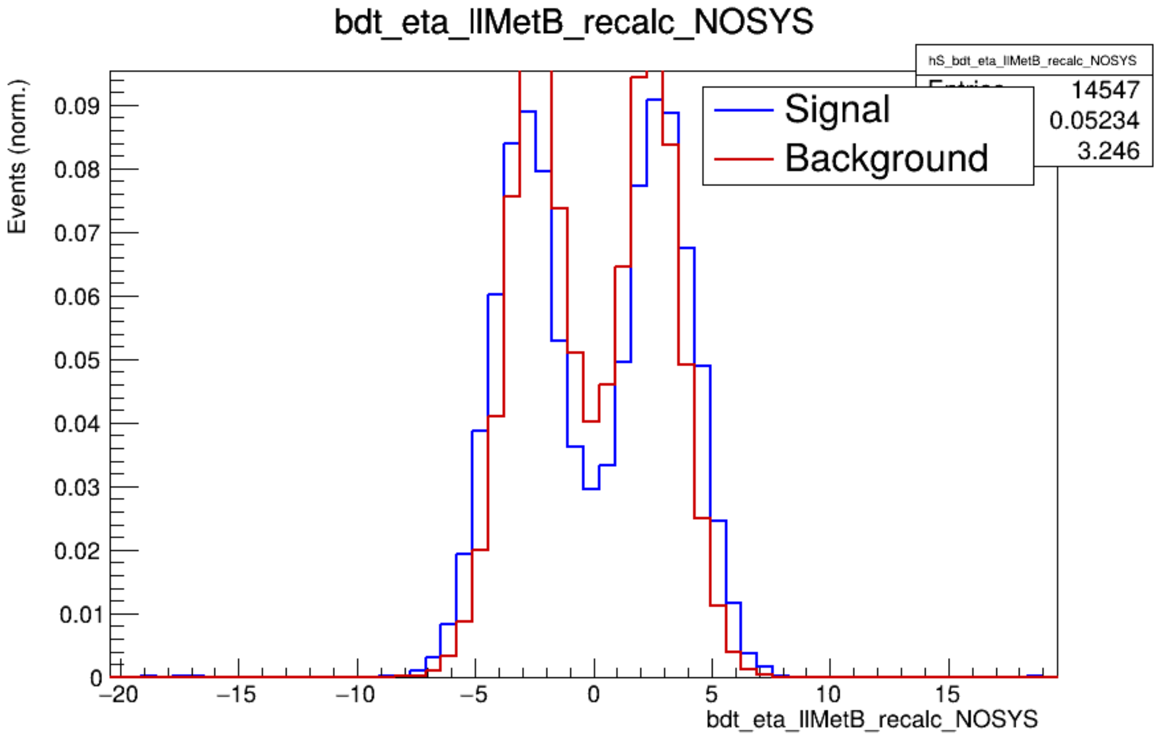


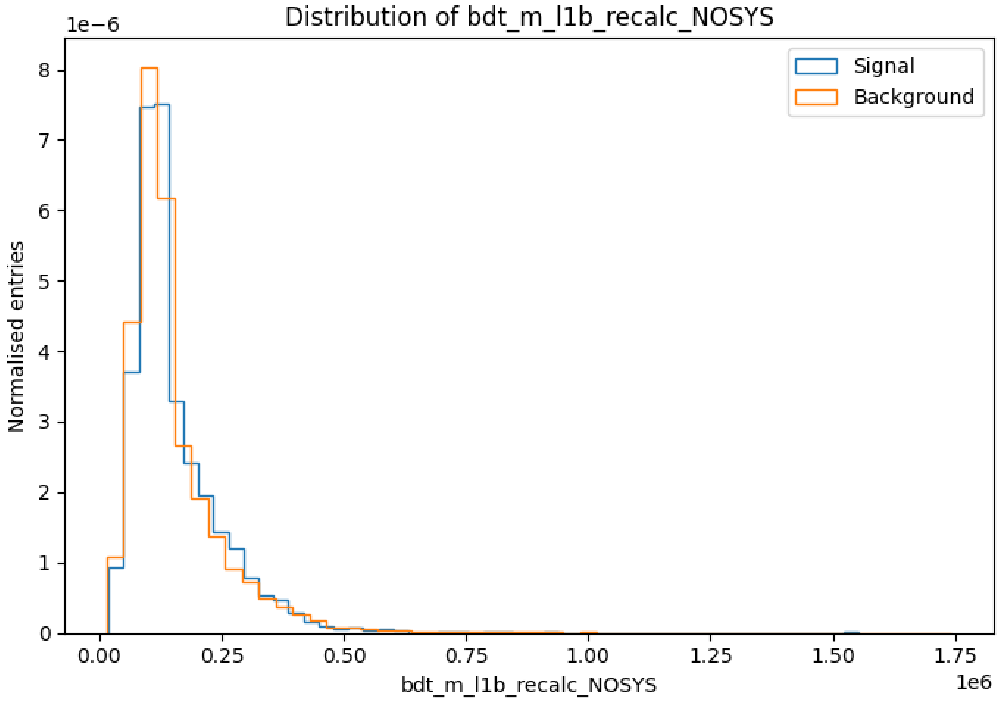
bdt_delta_pT_ll_MET_recalc_NOSYS
hS_bdt_delta_pT_ll_MET_recalc_NOSYS
6587
4.096e+04
3.67e+04
Signal
Background
Events (norm.)
×10^3
bdt_delta_pT_ll_MET_recalc_NOSYS

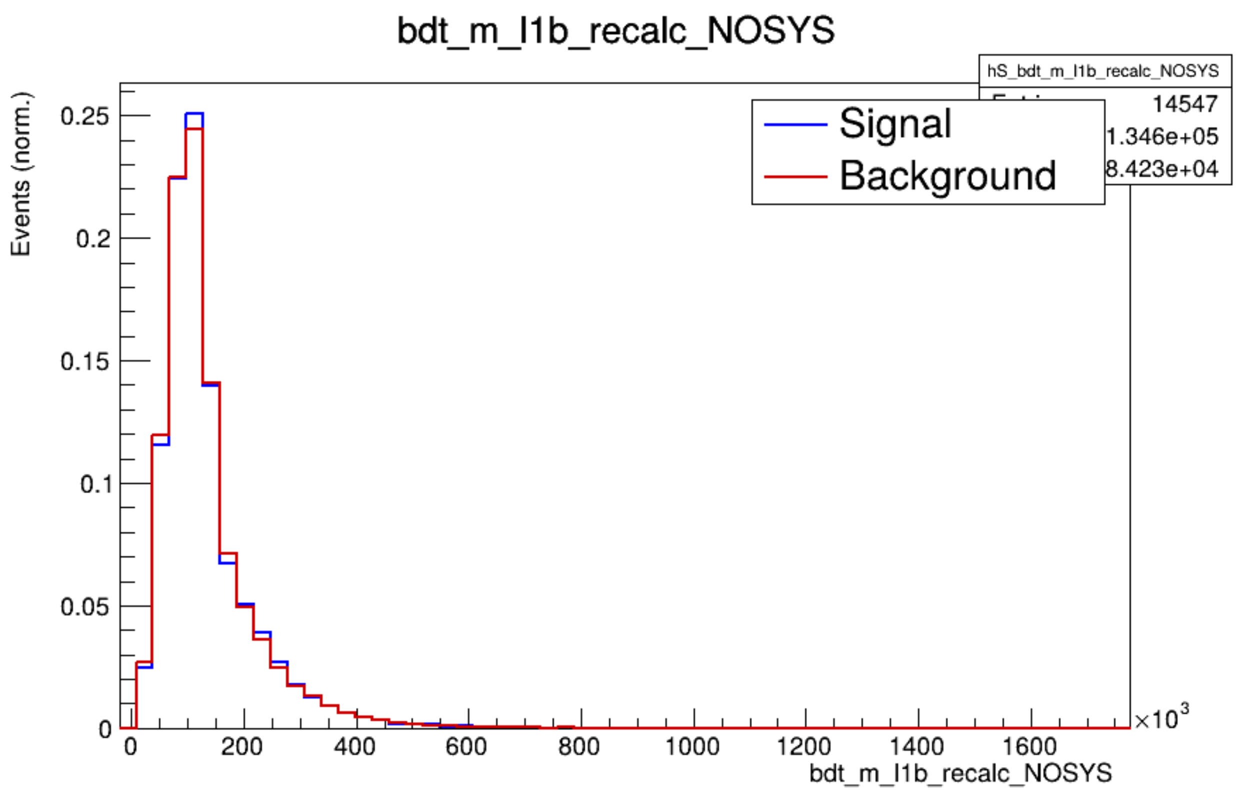

Distribution of bdt_delta_pT_llb_MET_recalc_NOSYS
1e−5
Signal
Background
Normalised entries
bdt_delta_pT_llb_MET_recalc_NOSYS

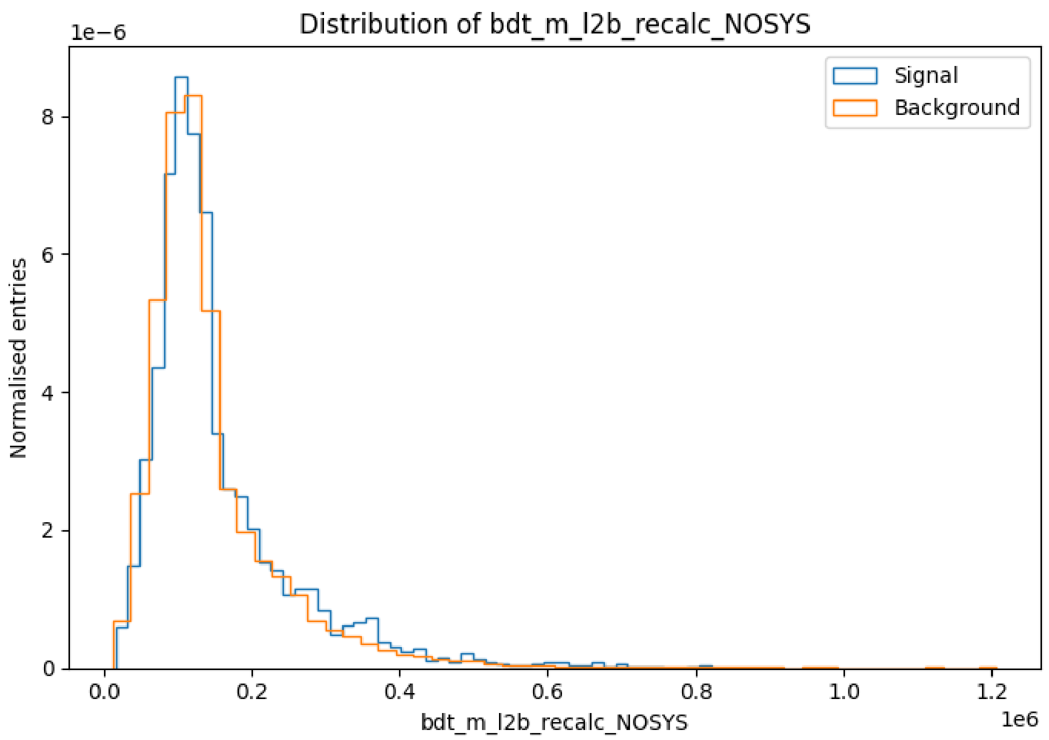

bdt_delta_pT_llb_MET_recalc_NOSYS
hS_bdt_delta_pT_llb_MET_recalc_NOSYS
7321
1.418e+04
1.98e+04
Signal
Background
Events (norm.)
×10³
bdt_delta_pT_llb_MET_recalc_NOSYS

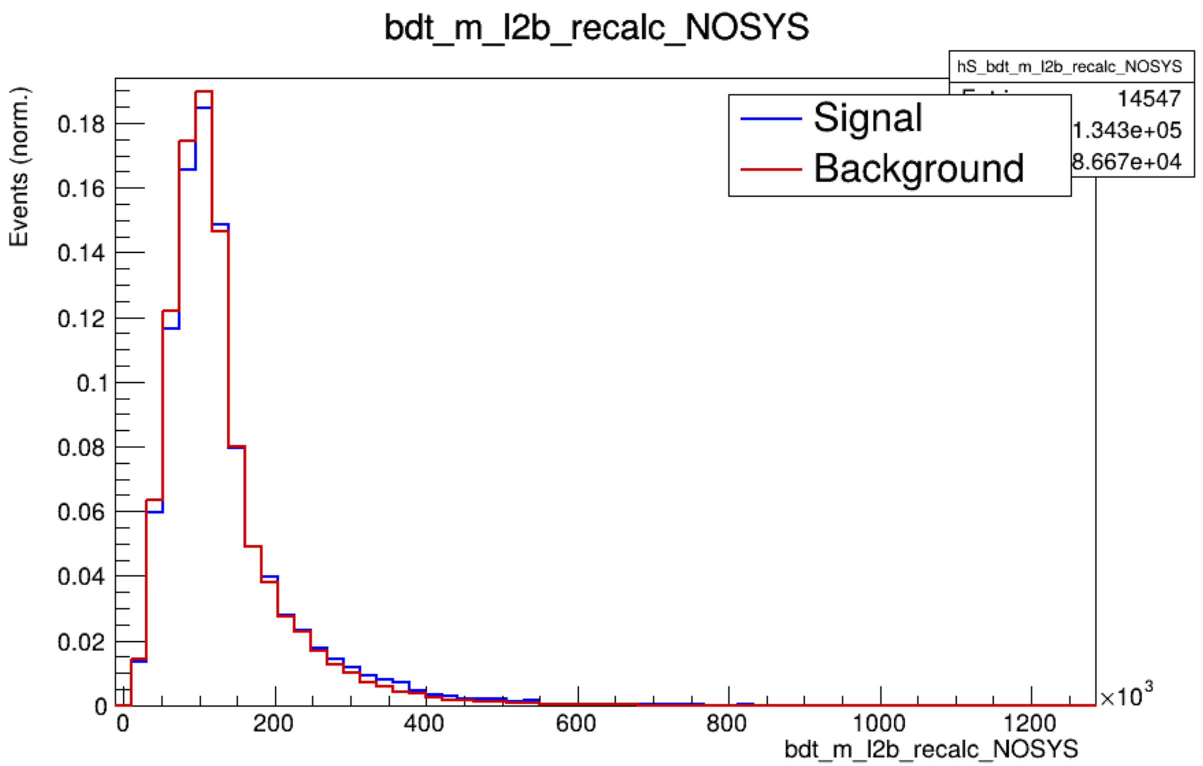

Distribution of bdt_eta_llMetB_recalc_NOSYS
Signal
Background
Normalised entries
bdt_eta_llMetB_recalc_NOSYS

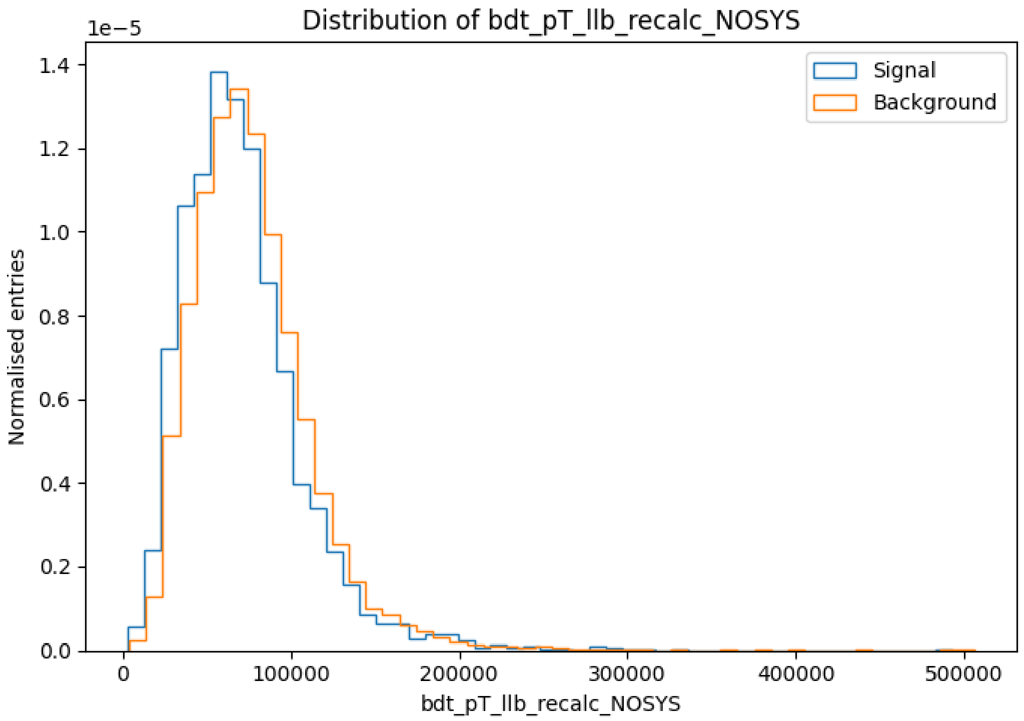

bdt_eta_llMetB_recalc_NOSYS
hS_bdt_eta_llMetB_recalc_NOSYS
14547
0.05234
3.246
Signal
Background
Events (norm.)
bdt_eta_llMetB_recalc_NOSYS

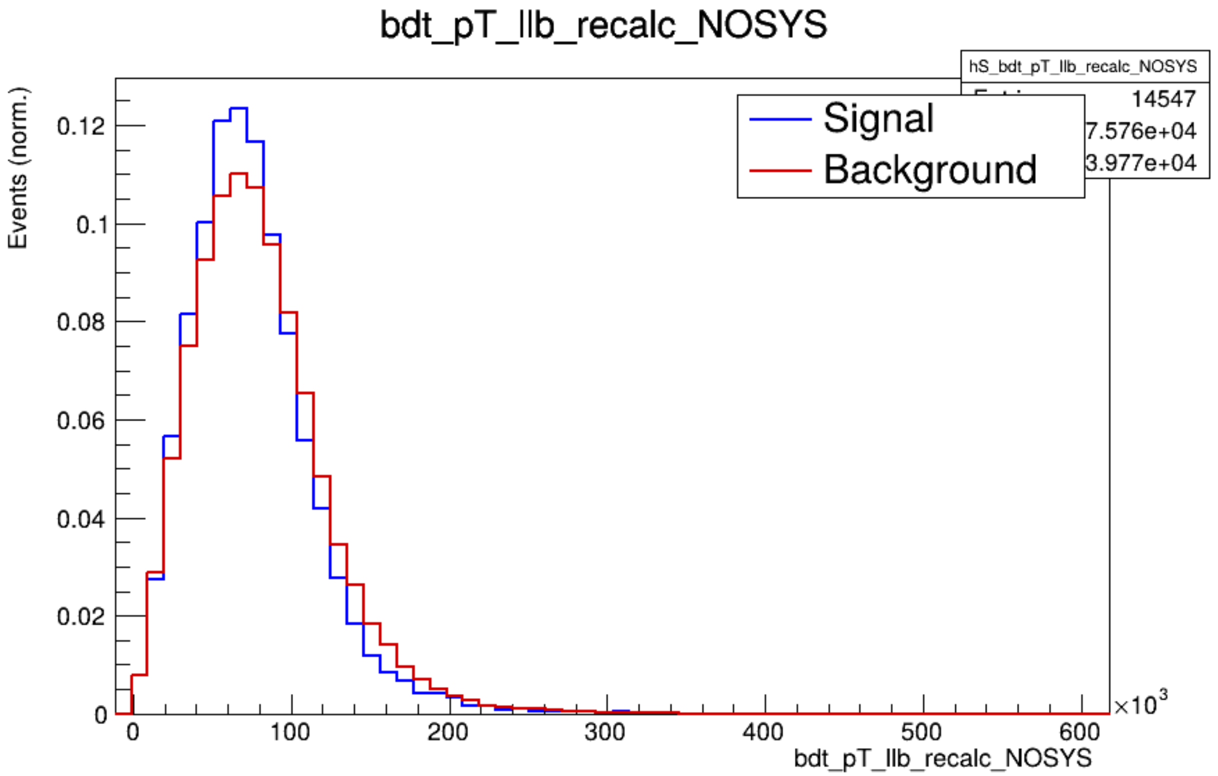

Distribution of bdt_m_l1b_recalc_NOSYS
1e−6
Signal
Background
Normalised entries
bdt_m_l1b_recalc_NOSYS
1e6

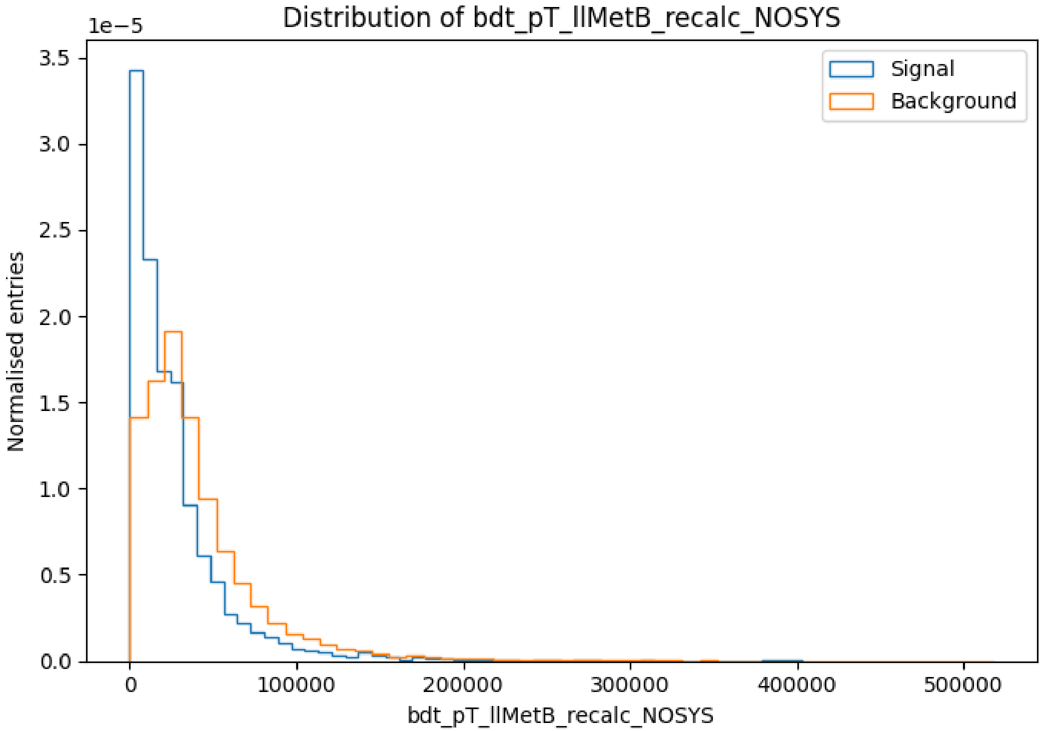

bdt_m_l1b_recalc_NOSYS
hS_bdt_m_l1b_recalc_NOSYS
14547
1.346e+05
8.423e+04
Signal
Background
Events (norm.)
×10^3
bdt_m_l1b_recalc_NOSYS


Distribution of bdt_m_l2b_recalc_NOSYS
1e−6
Signal
Background
Normalised entries
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6
bdt_m_l2b_recalc_NOSYS

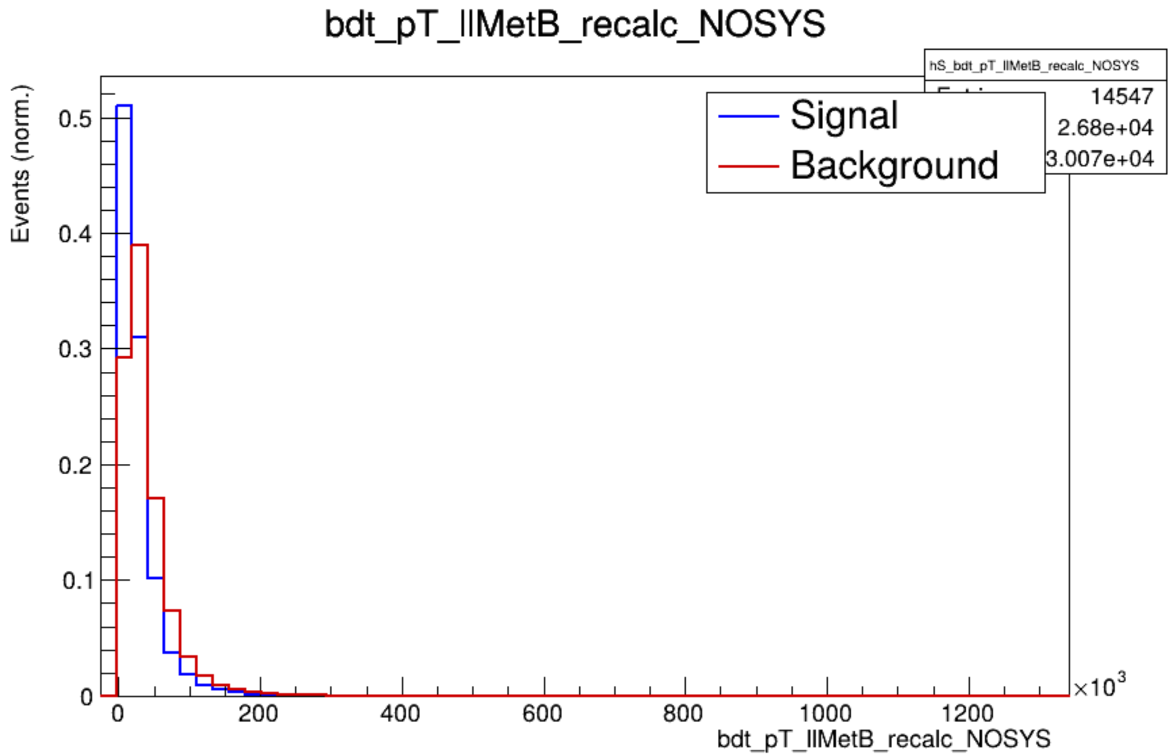

bdt_m_l2b_recalc_NOSYS
hS_bdt_m_l2b_recalc_NOSYS
14547
1.343e+05
8.667e+04
Signal
Background
Events (norm.)
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0
200
400
600
800
1000
1200
$\times10^{3}$
bdt_m_l2b_recalc_NOSYS


Distribution of bdt_pT_llb_recalc_NOSYS
1e−5
Signal
Background
Normalised entries
bdt_pT_llb_recalc_NOSYS


bdt_pT_llb_recalc_NOSYS
hS_bdt_pT_llb_recalc_NOSYS
14547
7.576e+04
3.977e+04
Signal
Background
Events (norm.)
×10^3
bdt_pT_llb_recalc_NOSYS



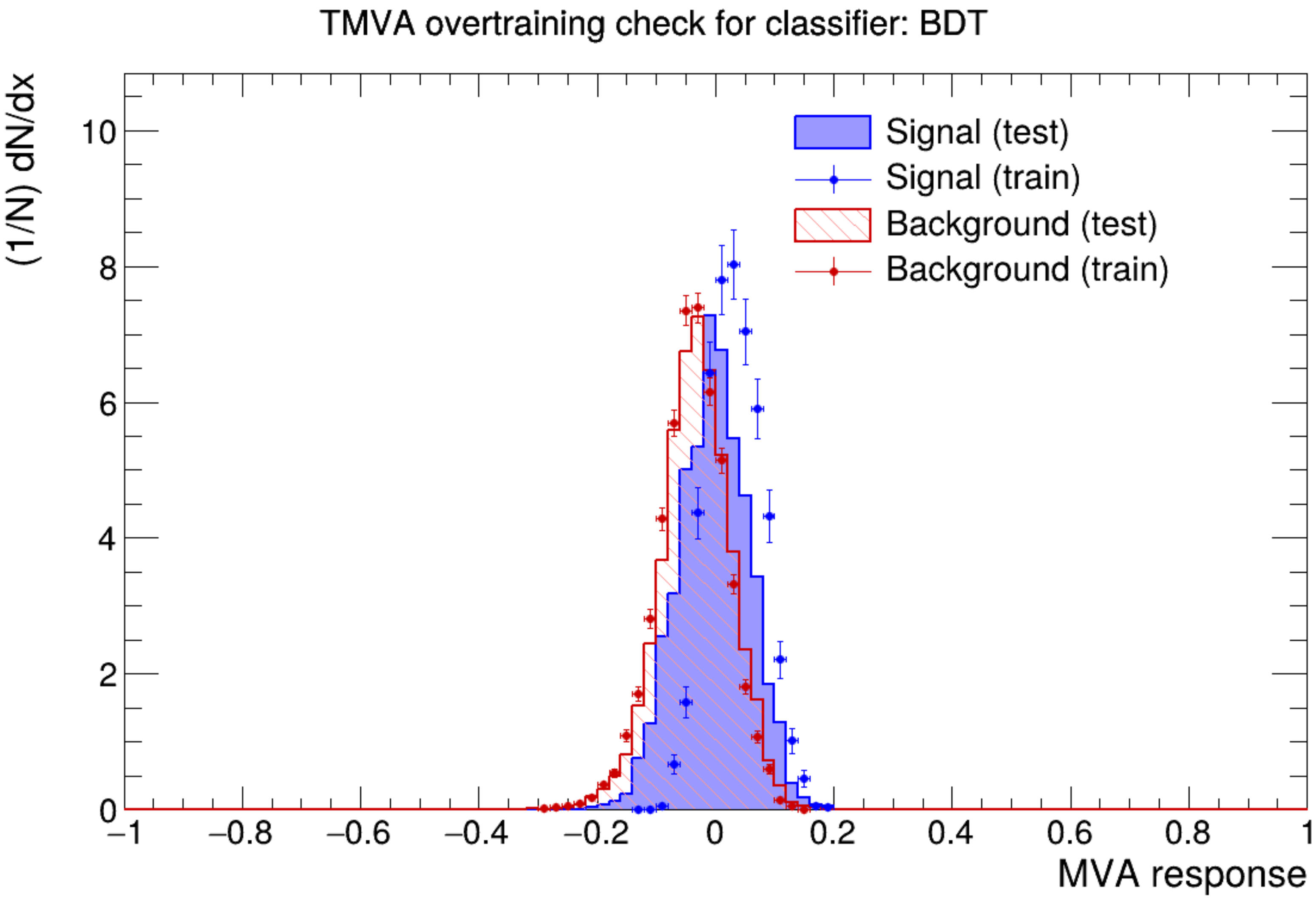
Distribution of bdt_pT_llMetB_recalc_NOSYS
1e−5
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Normalised entries
Signal
Background
0
100000
200000
300000
400000
500000
bdt_pT_llMetB_recalc_NOSYS



bdt_pT_llMetB_recalc_NOSYS
hS_bdt_pT_llMetB_recalc_NOSYS
14547
2.68e+04
3.007e+04
Signal
Background
Events (norm.)
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
×10^3
bdt_pT_llMetB_recalc_NOSYS

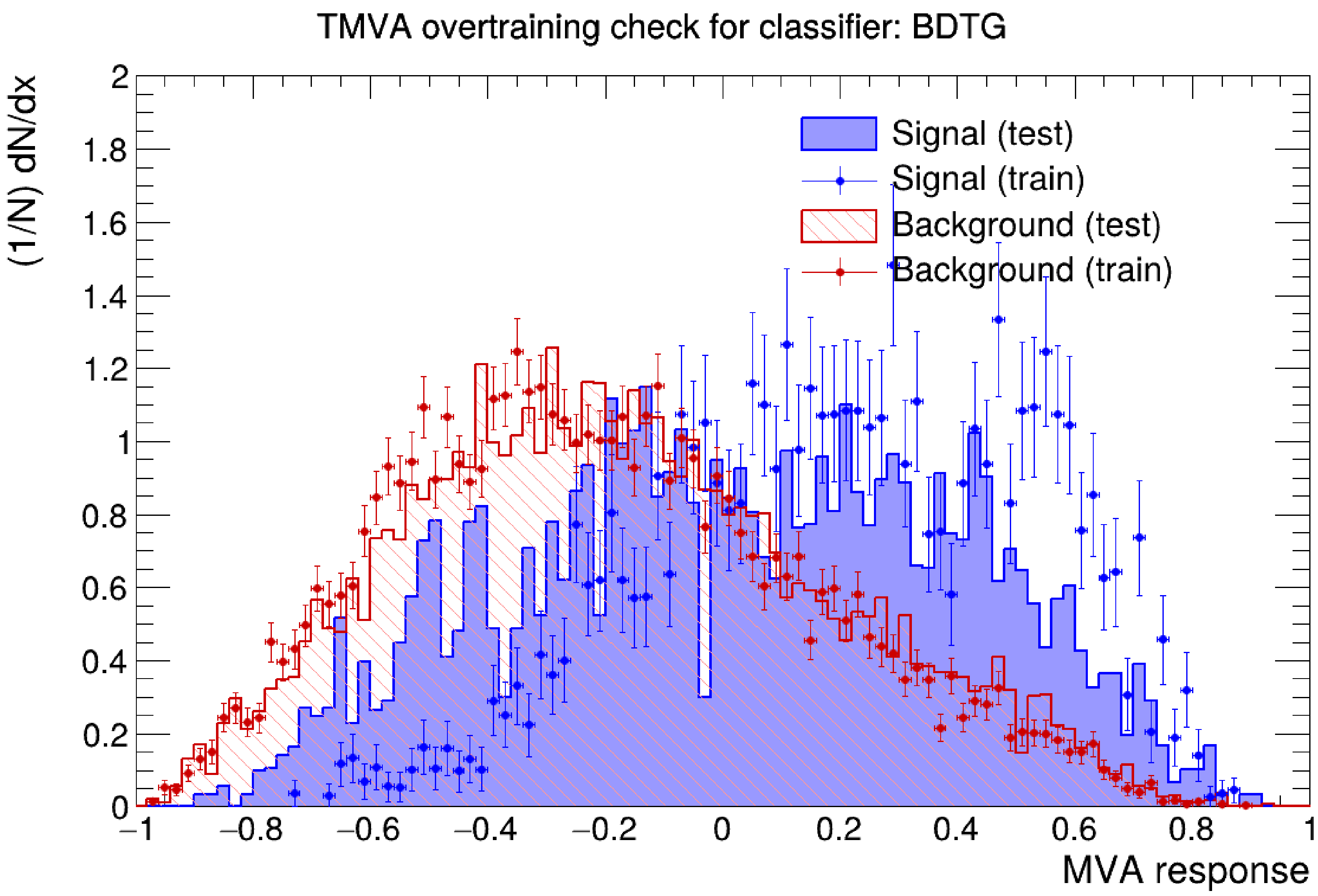
Results of TMVA on new data 16.12.2025

BDT

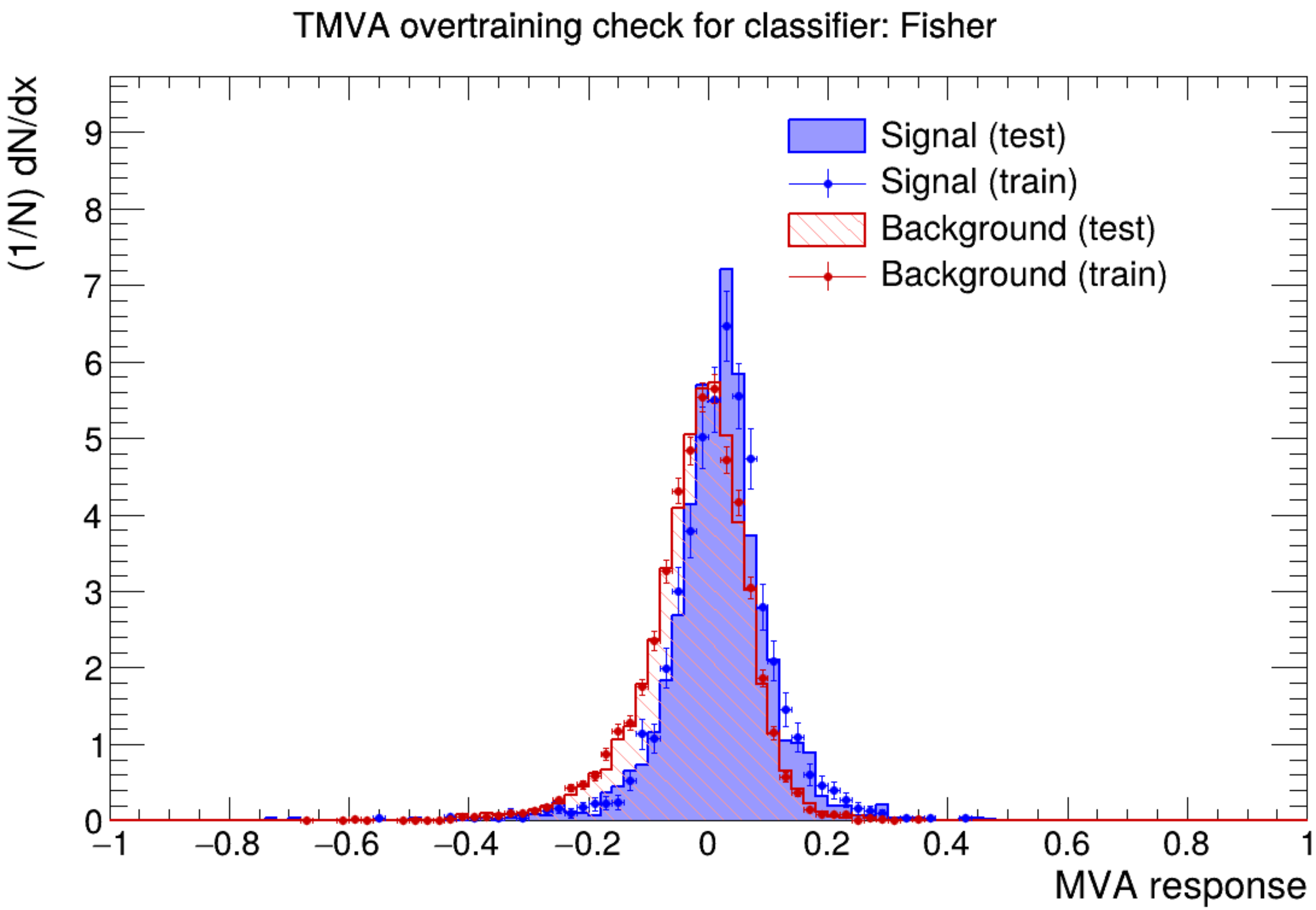


TMVA overtraining check for classifier: BDT
(1/N) dN/dx
10
8
6
4
2
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
MVA response
Signal (test)
Signal (train)
Background (test)
Background (train)

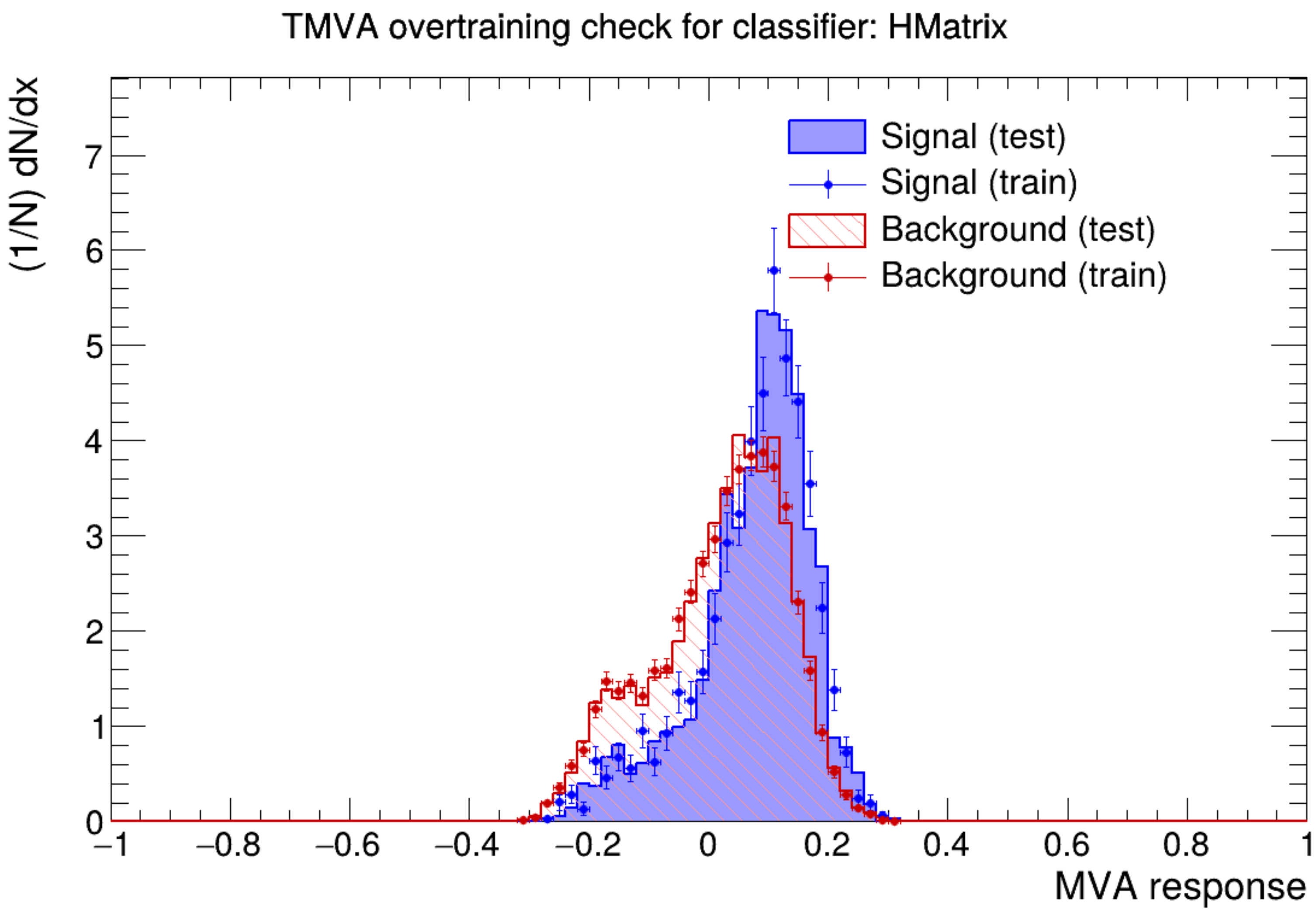
BDTG


TMVA overtraining check for classifier: BDTG
(1/N) dN/dx
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
MVA response
Signal (test)
Signal (train)
Background (test)
Background (train)

Fisher discriminant TMVA

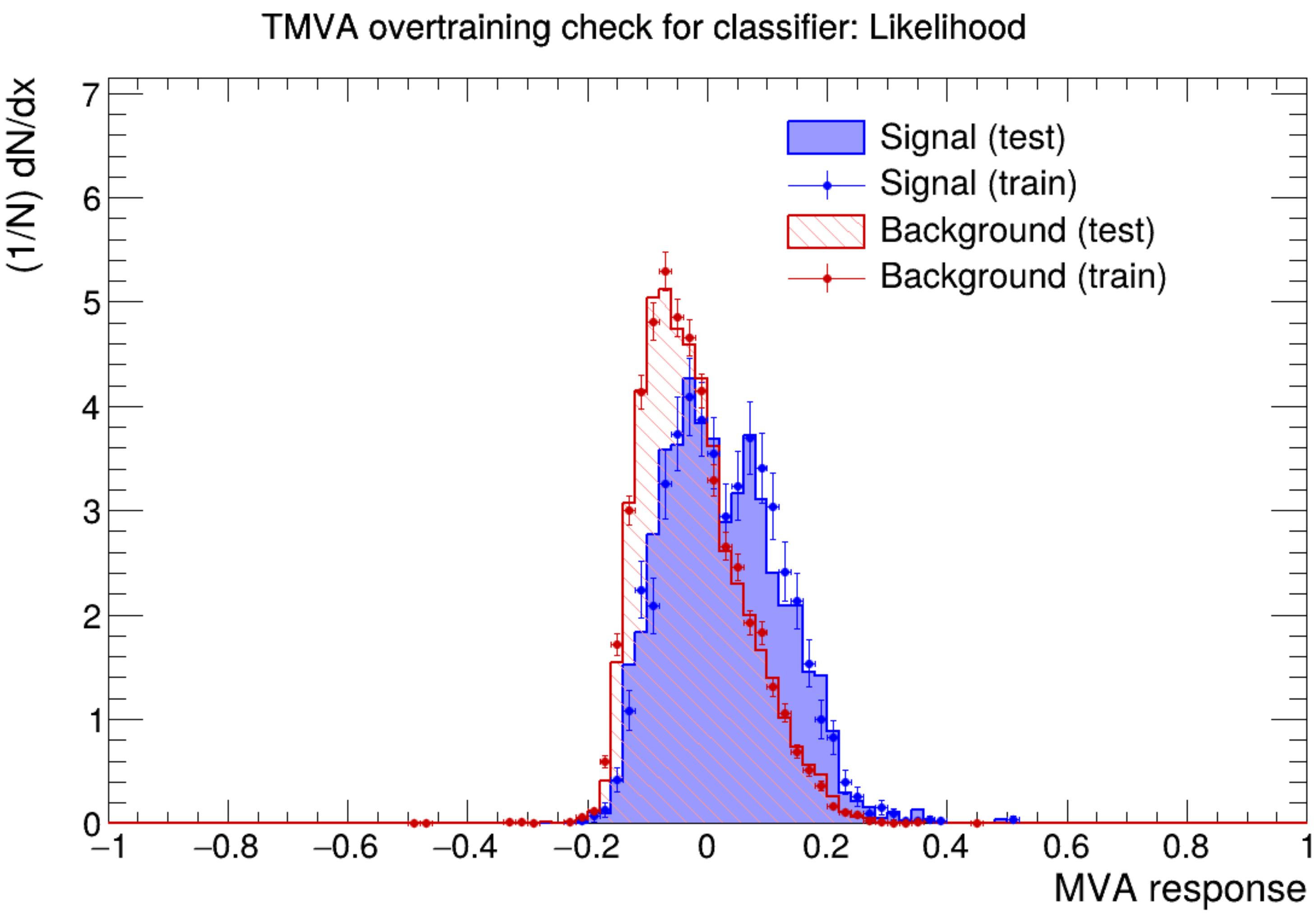


H-Matrix

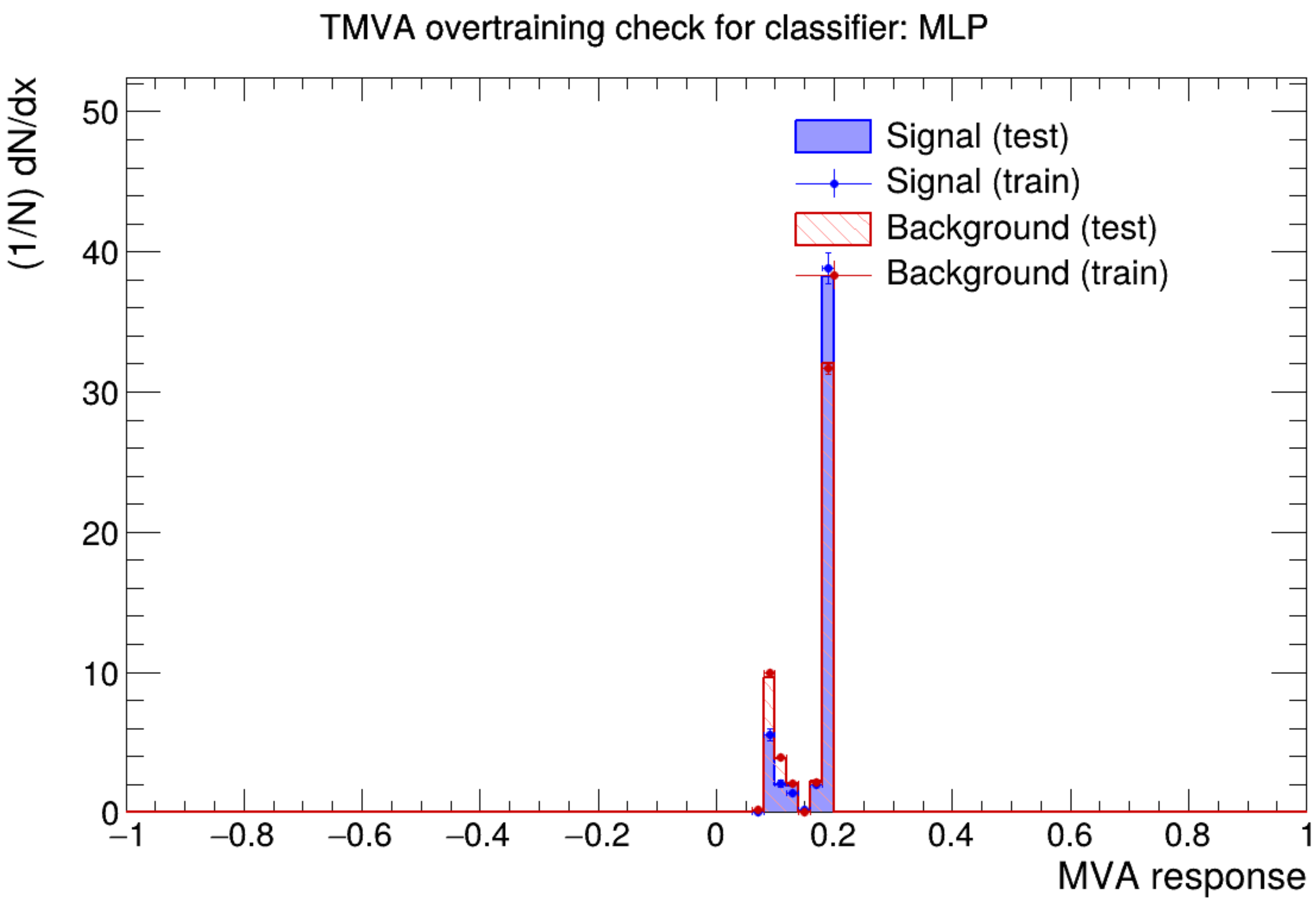


TMVA overtraining check for classifier: HMatrix
(1/N) dN/dx
0
1
2
3
4
5
6
7
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
MVA response
Signal (test)
Signal (train)
Background (test)
Background (train)

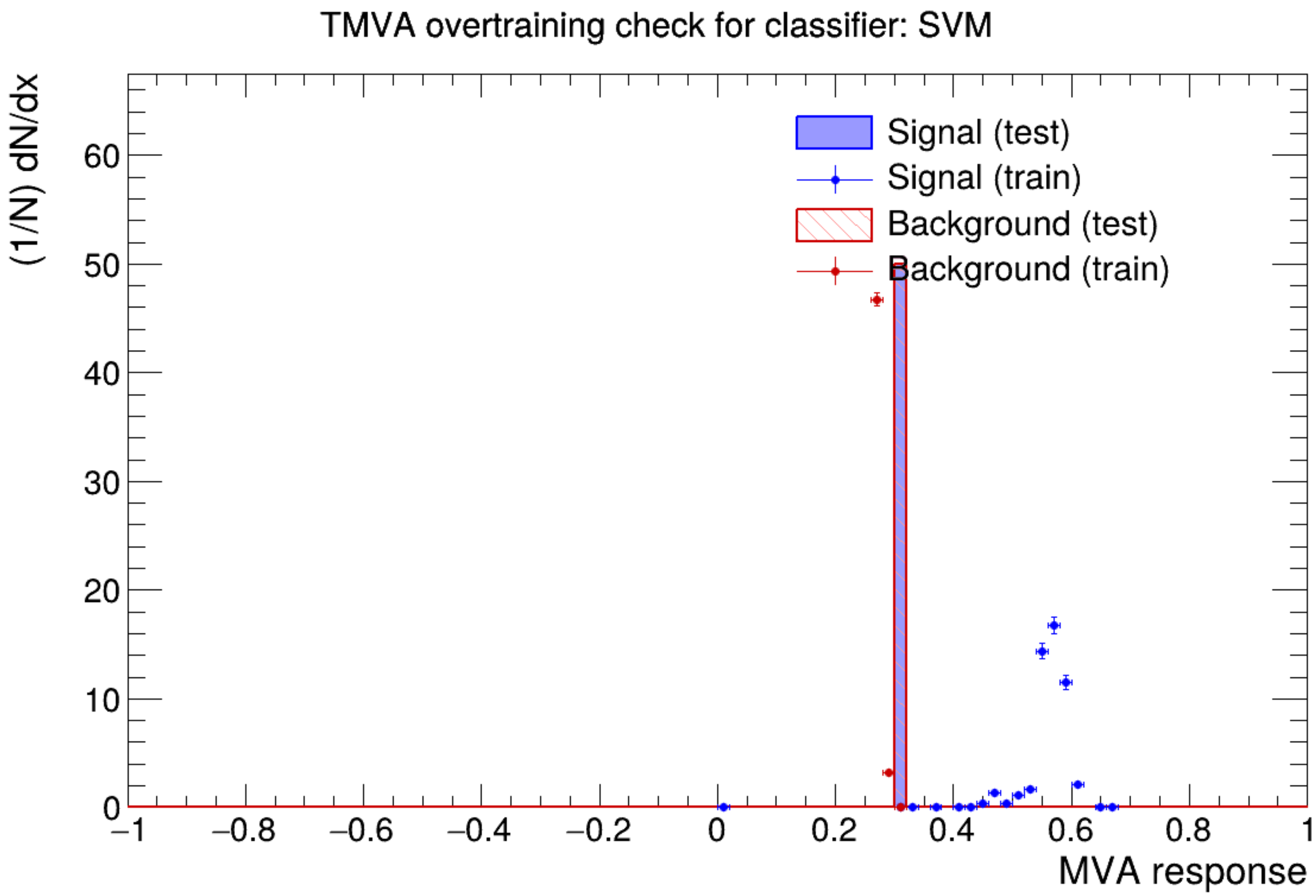
Likelihood Estimate TMVA



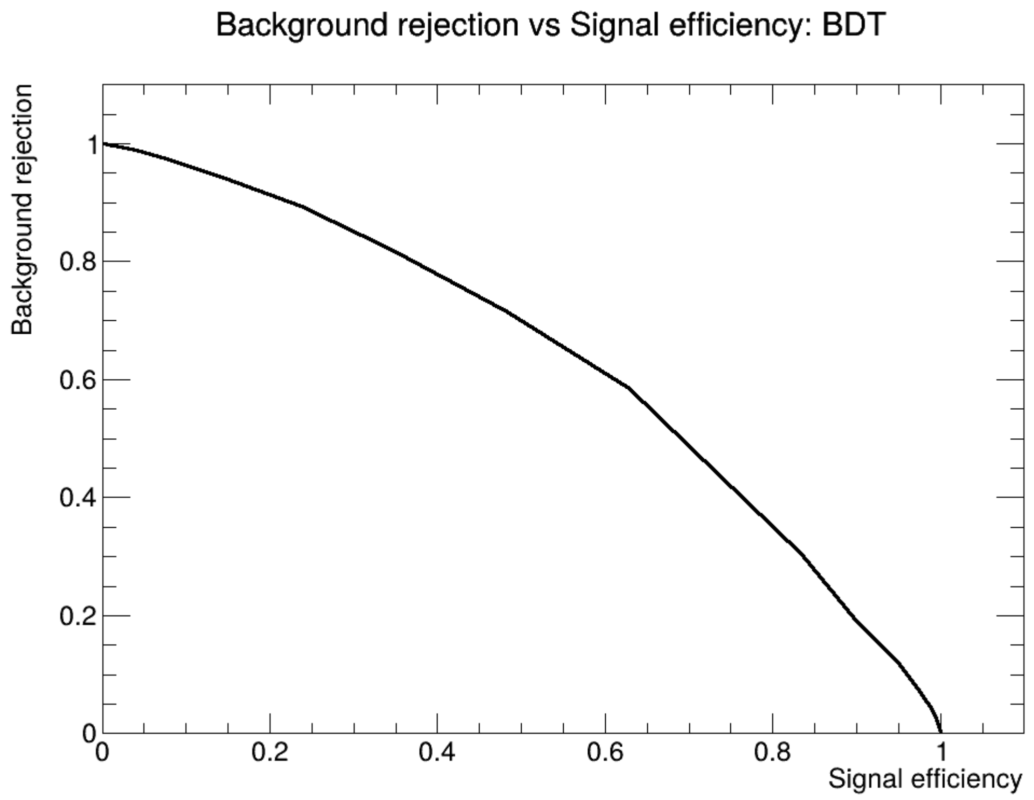
MLP

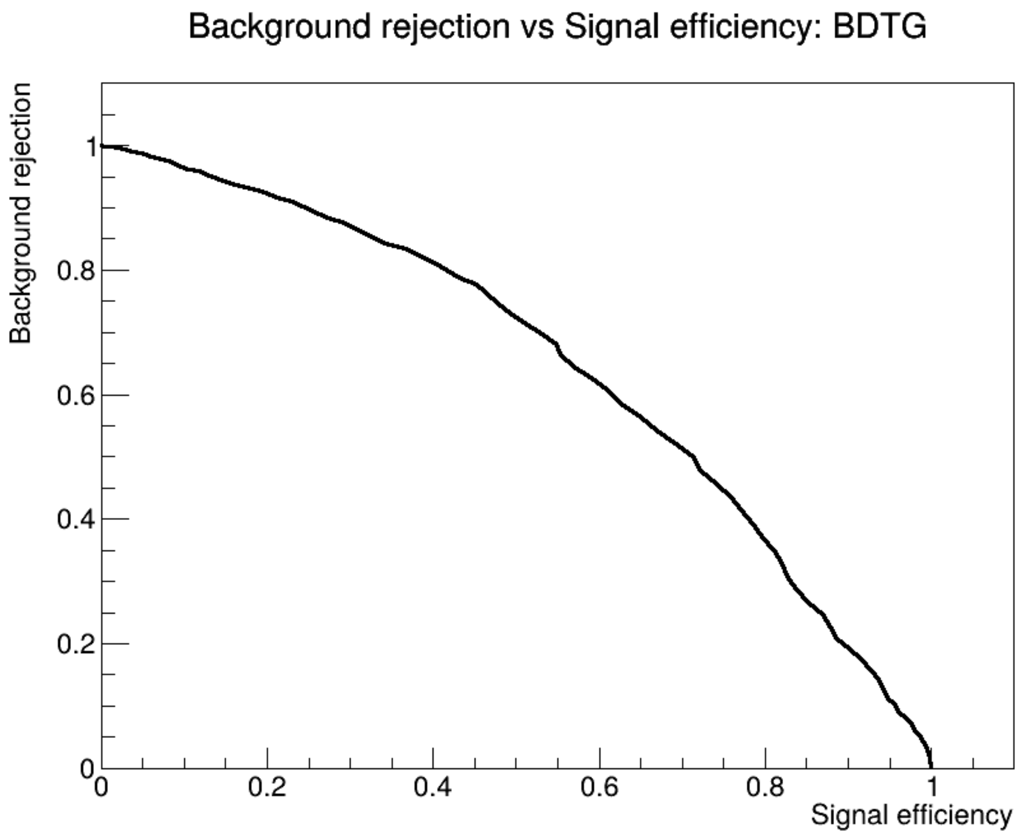

TMVA overtraining check for classifier: MLP
(1/N) dN/dx
50
40
30
20
10
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
MVA response
Signal (test)
Signal (train)
Background (test)
Background (train)

SVM

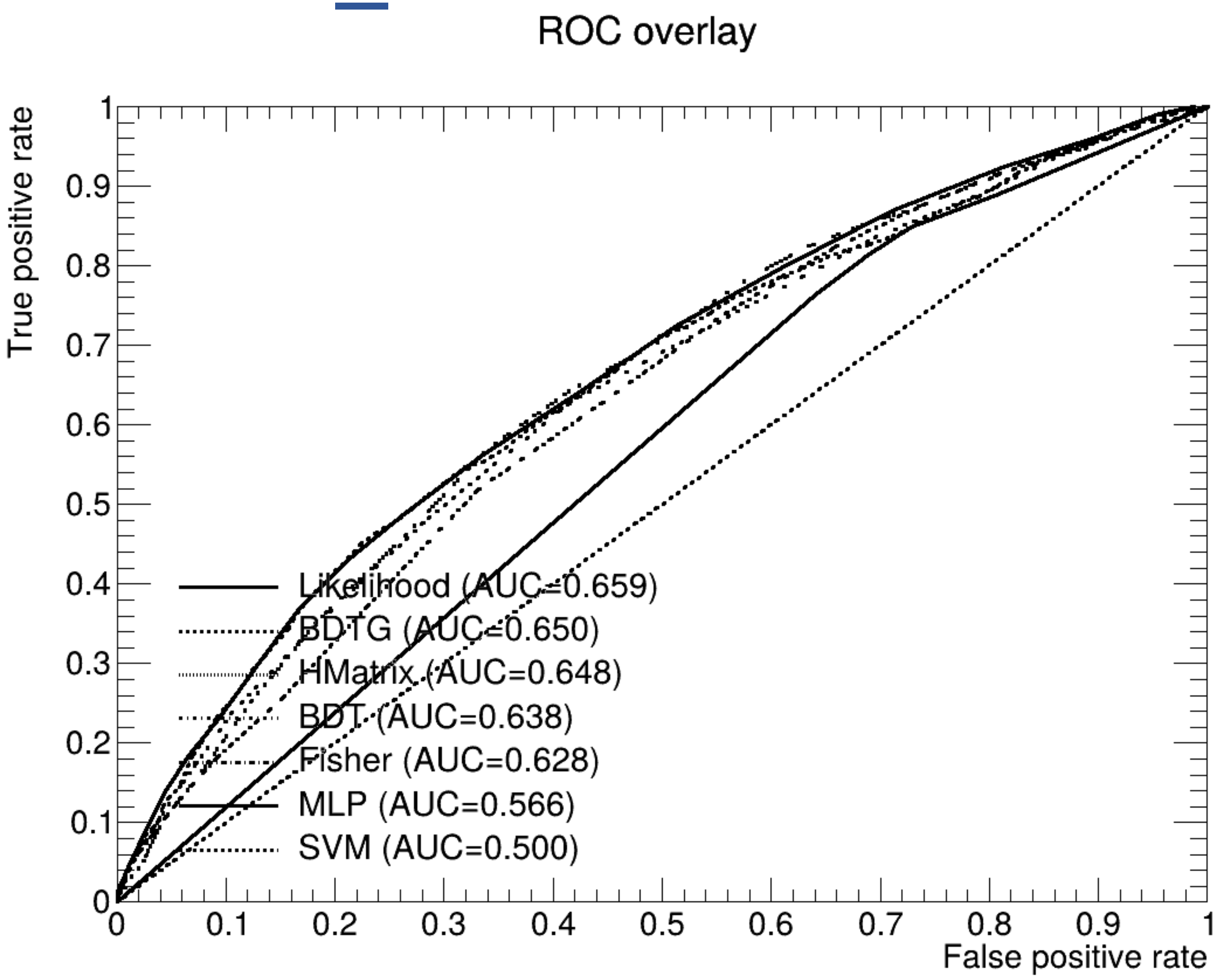

TMVA overtraining check for classifier: SVM
(1/N) dN/dx
60
50
40
30
20
10
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
MVA response
Signal (test)
Signal (train)
Background (test)
Background (train)

Background vs Signal efficiency BDT and BDTG





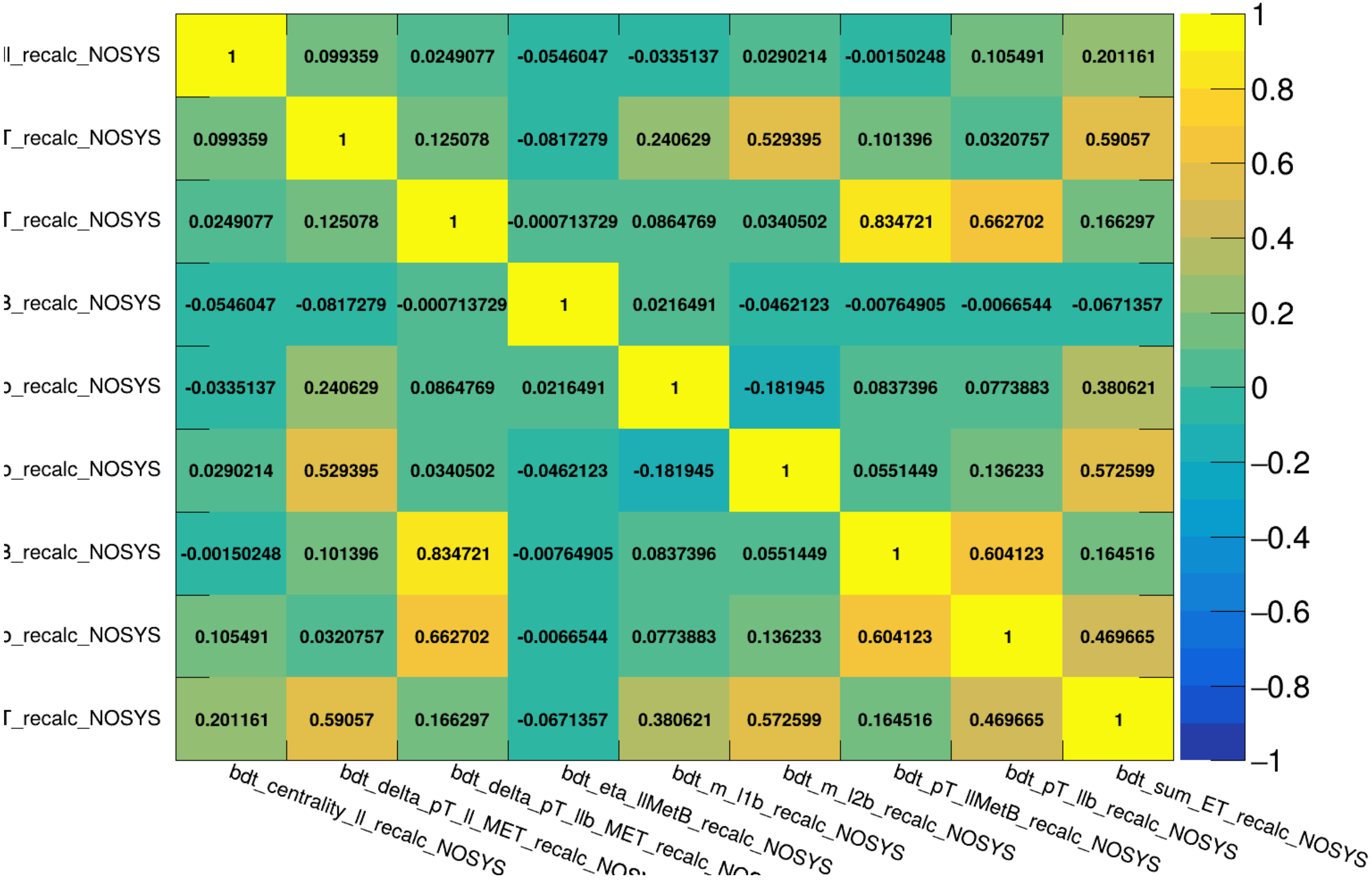
All_ROC TMVA


ROC overlay
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Likelihood (AUC=0.659)
BDTG (AUC=0.650)
HMatrix (AUC=0.648)
BDT (AUC=0.638)
Fisher (AUC=0.628)
MLP (AUC=0.566)
SVM (AUC=0.500)

Comparison of performance of algorithms

Algorithm	**AUC**
BDTG	**0.650342**
BDT	0.637717
Fisher	0.627763
HMatrix	0.648258
Likelihood	**0.658574**
MLP	0.566076
SVM	0.5

Correlation (Signal) â original inputs



National Research
Tomsk
State
University

Лаборатория
анализа данных
физики высоких энергий
Томского
государственного
университета

Thank you for your attention!!!