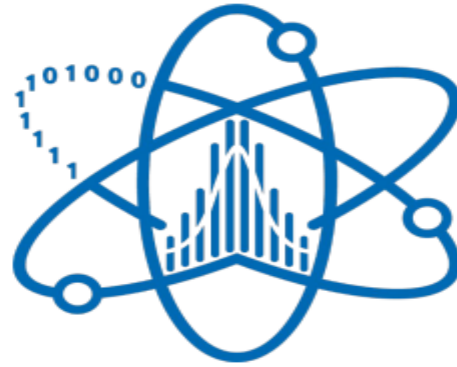




National Research
**Tomsk
State
University**



**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
9th December 2025**

Neda Firoz

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

- **Inputs (9 variables used):**

bdt_centralty_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) + $t\bar{W}$ (anti-top)

Background: $t\bar{t}$ (non-all-had)

- **Event weights:** auto-resolved to
 $\text{weight_mc_NOSYS} * \text{weight_pileup_NOSYS} * \text{globalTriggerEffSF_NOSYS}$.

- **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).

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

- 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)

- 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

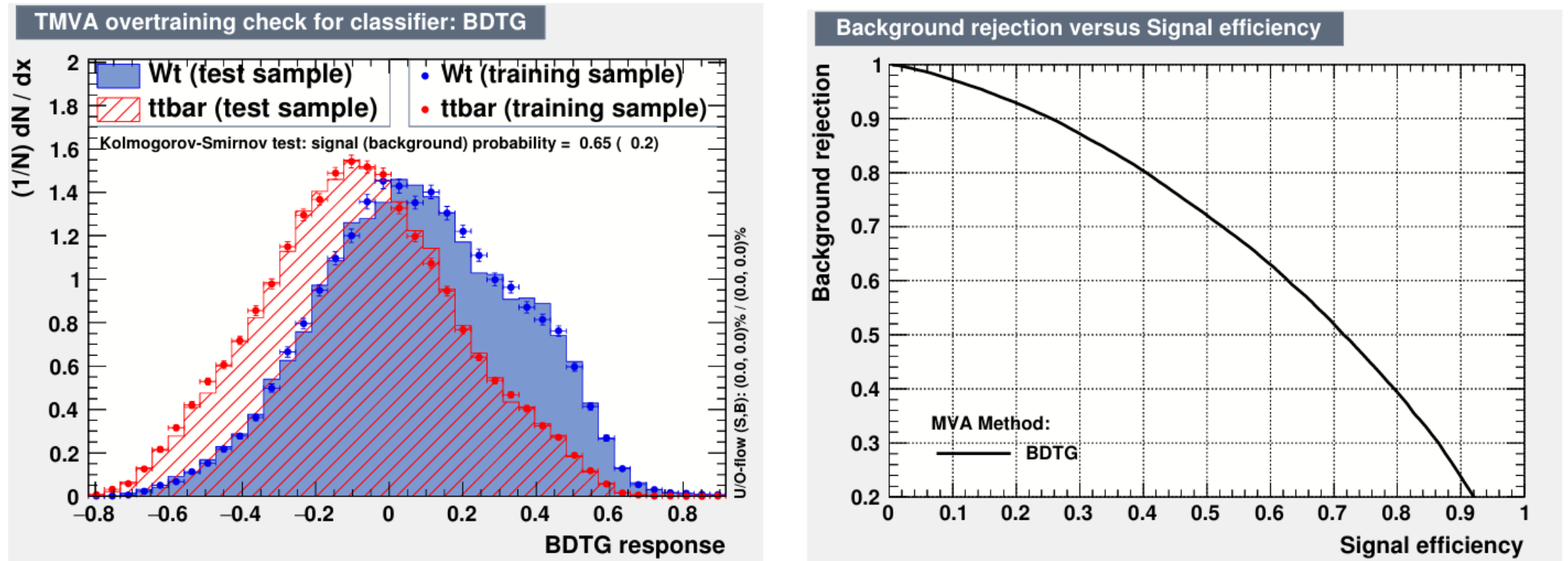
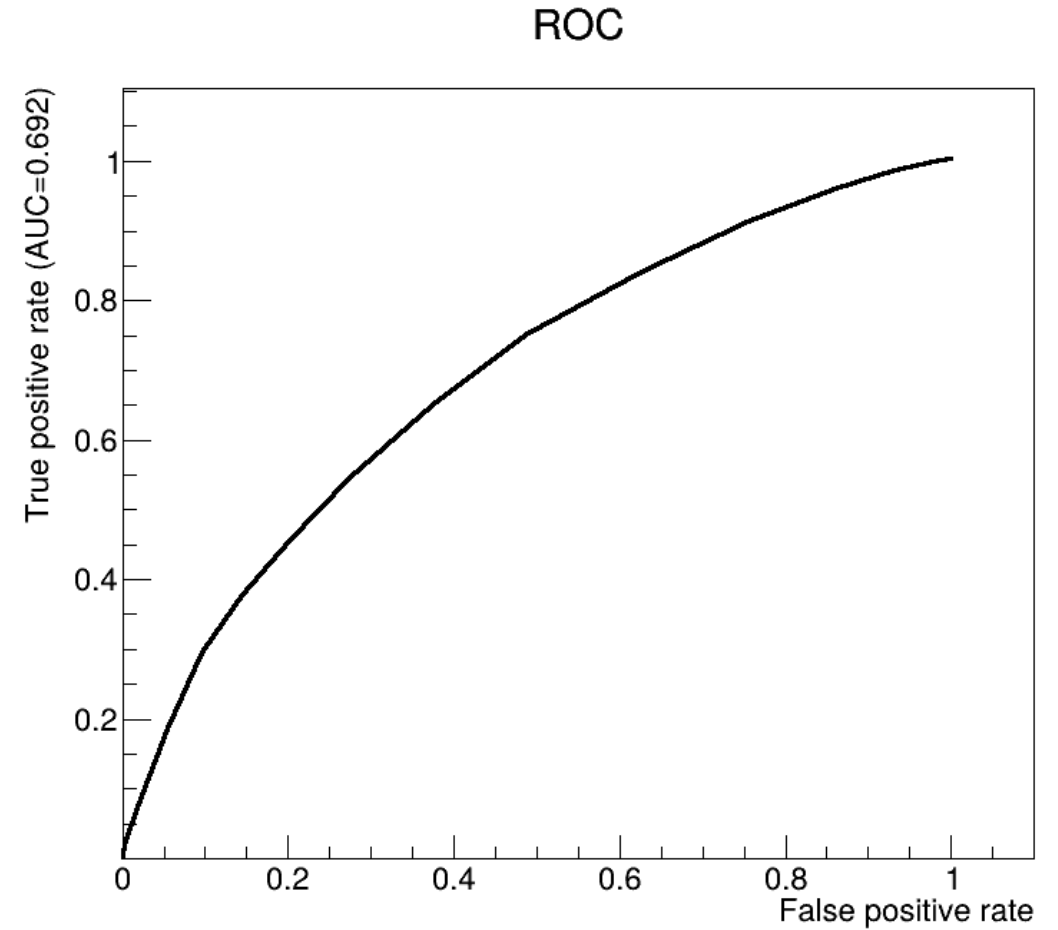
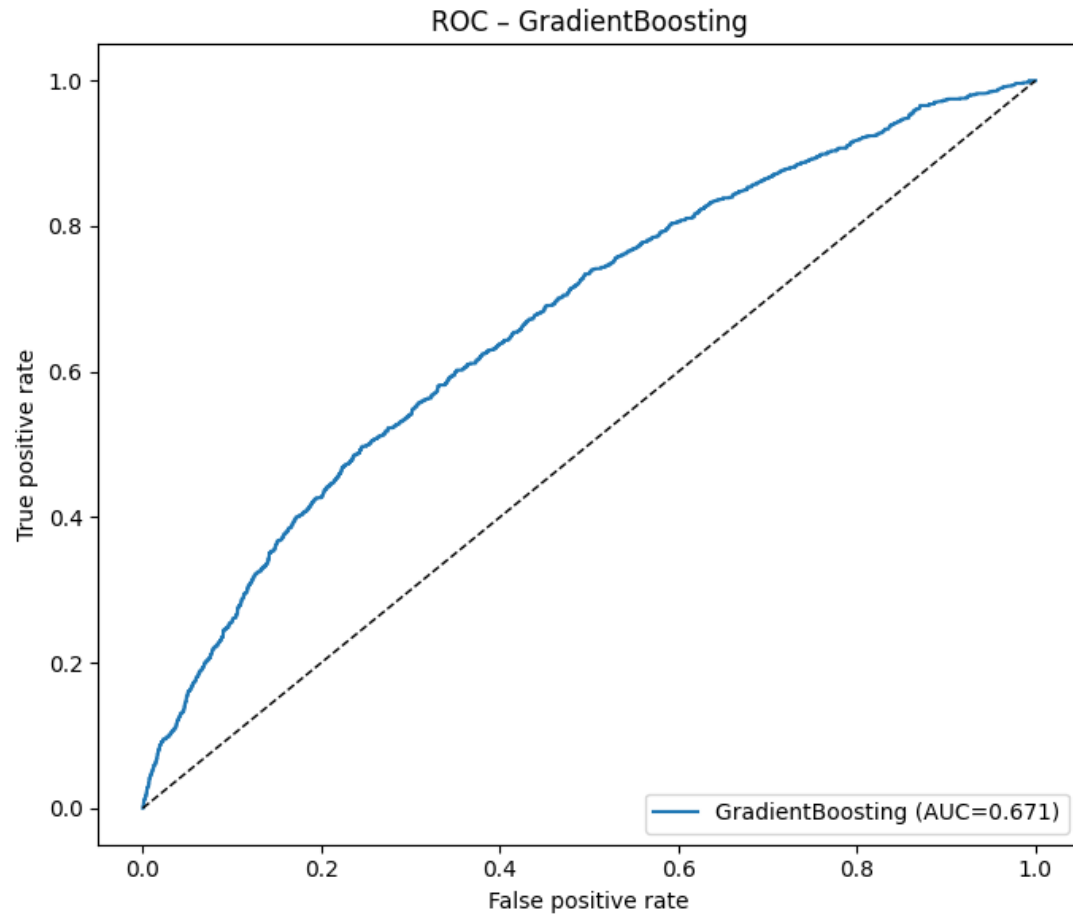


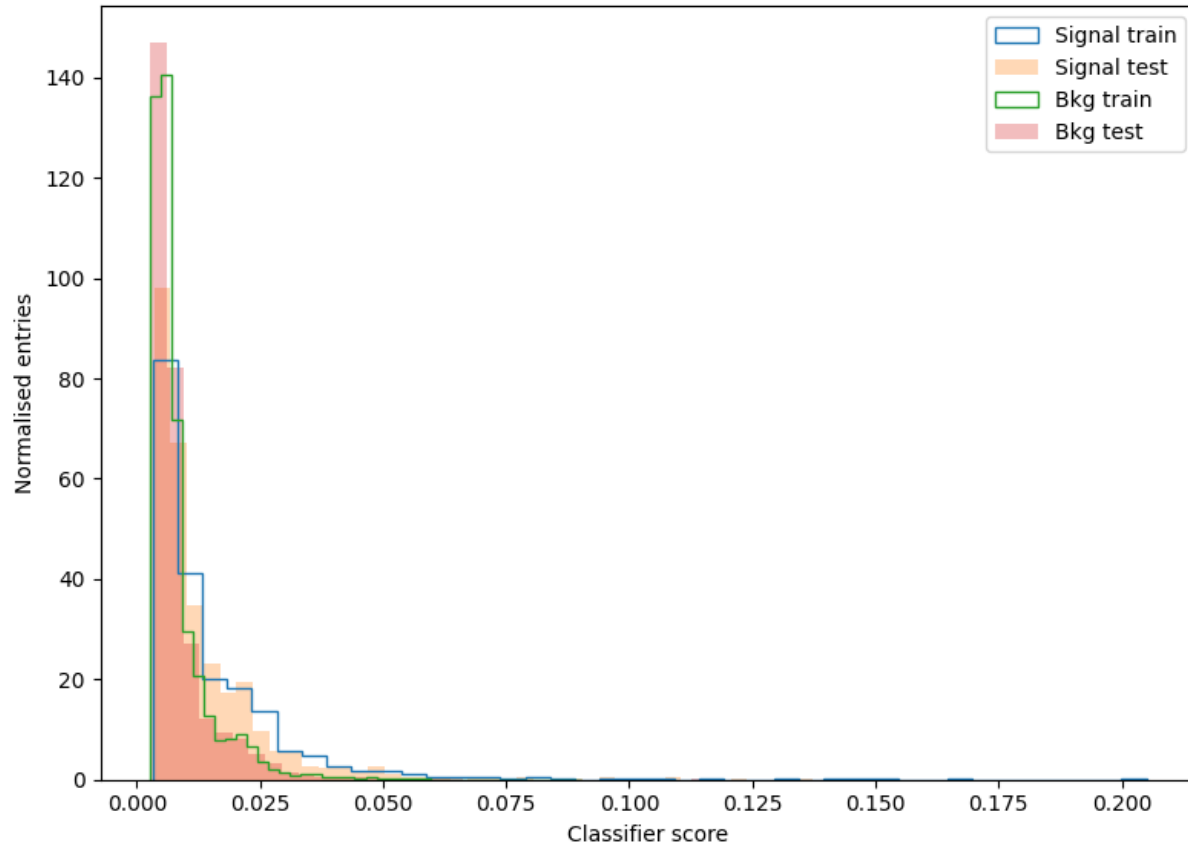
Figure 9: Comparison of test/training sample distributions and background rejection factor versus signal efficiency.

ROC of BDT from sklearn and TMVA

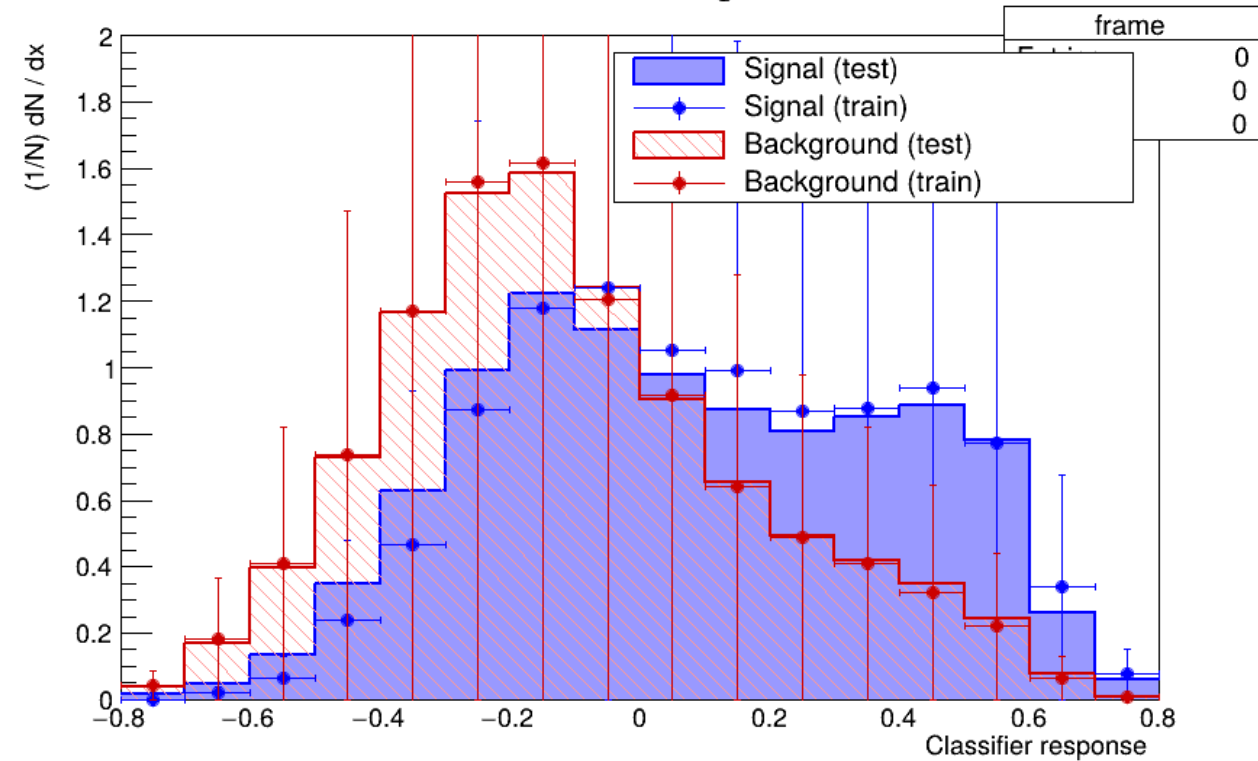


Gradient Boosting Algorithm from R and BDT from TMVA

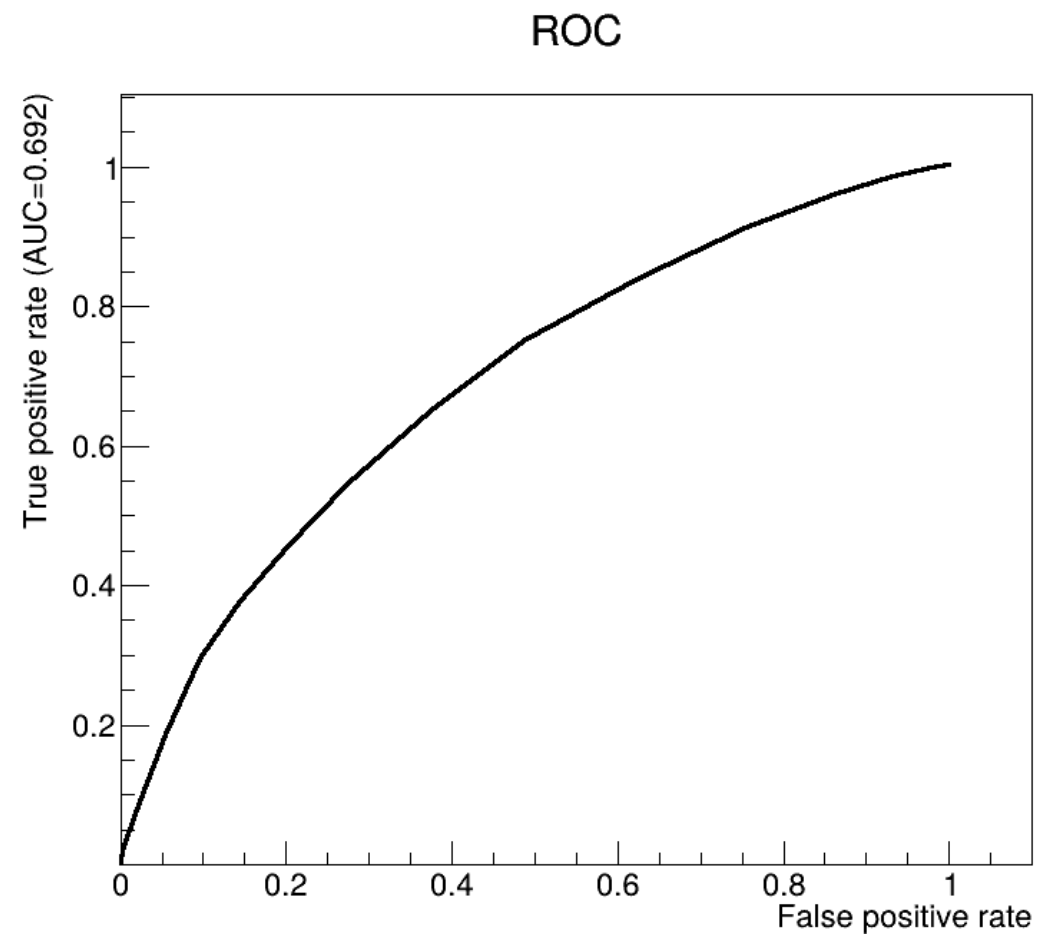
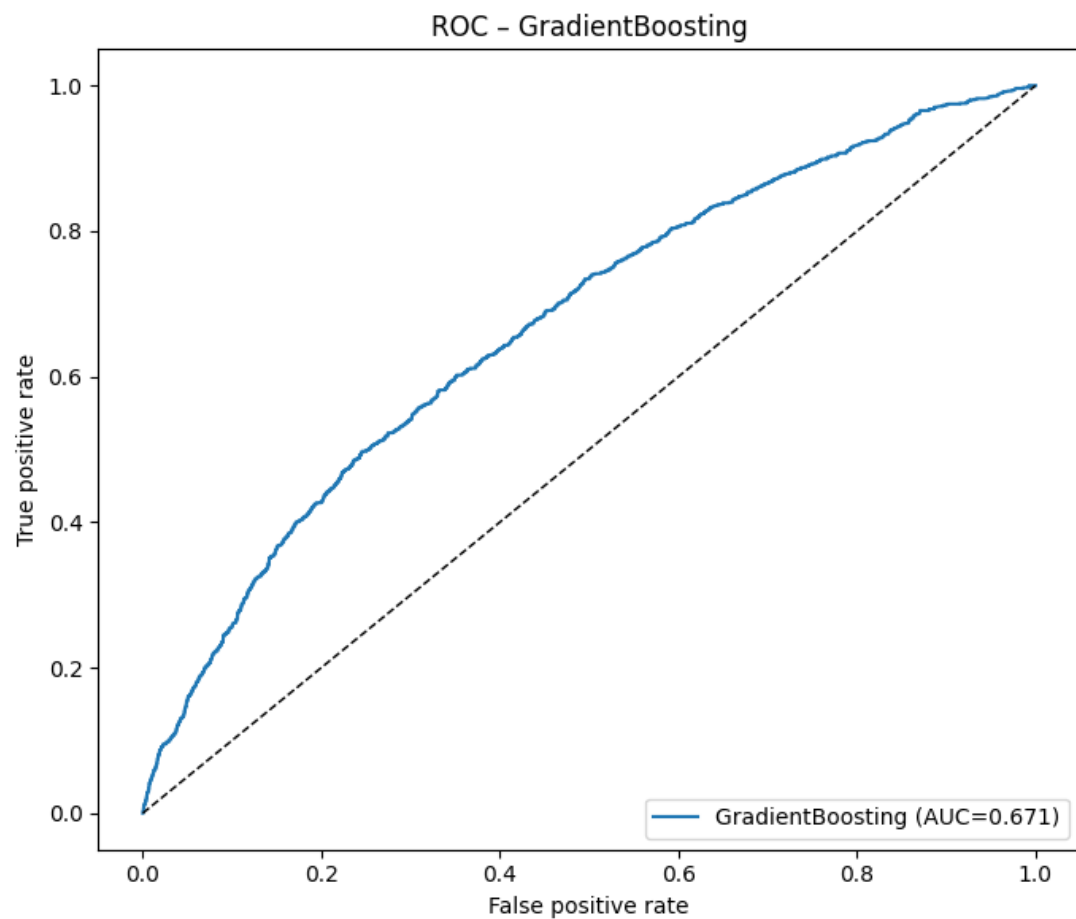
Overtraining check - GradientBoosting



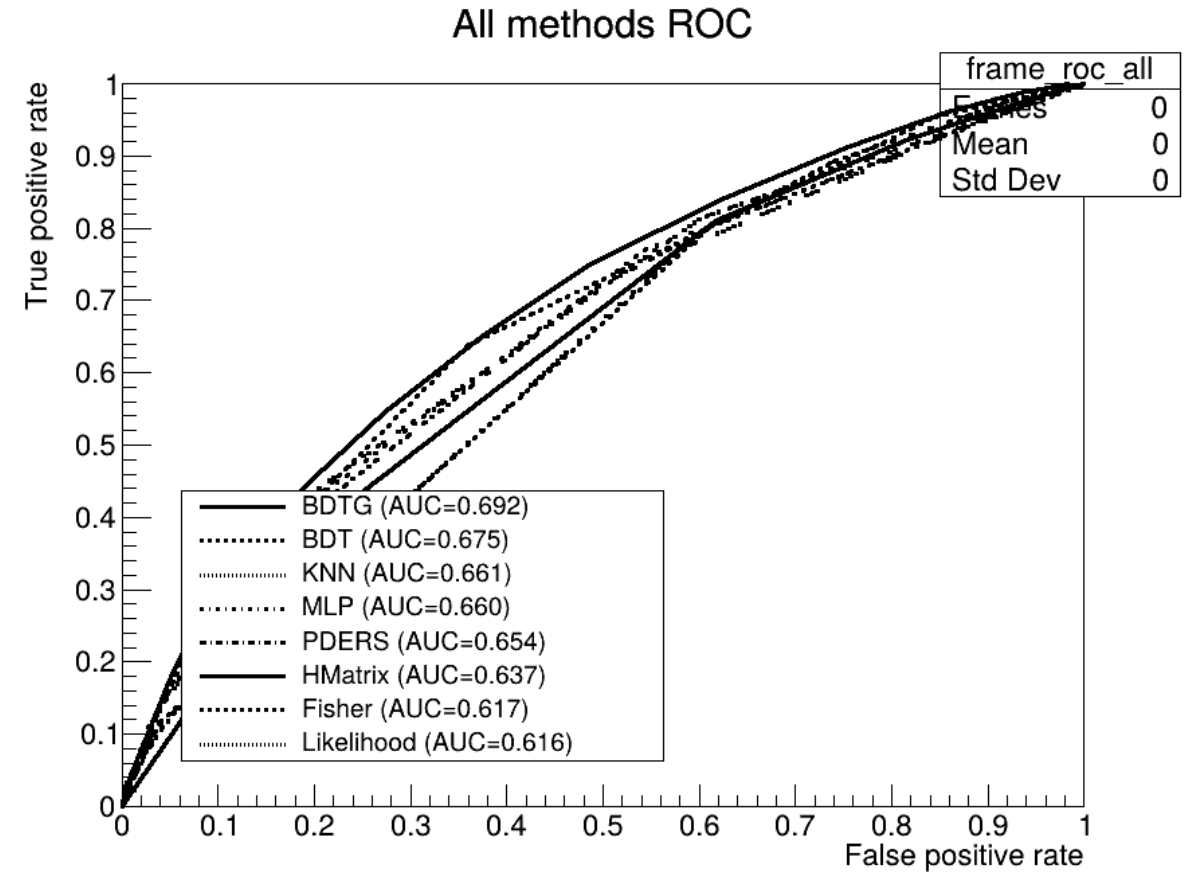
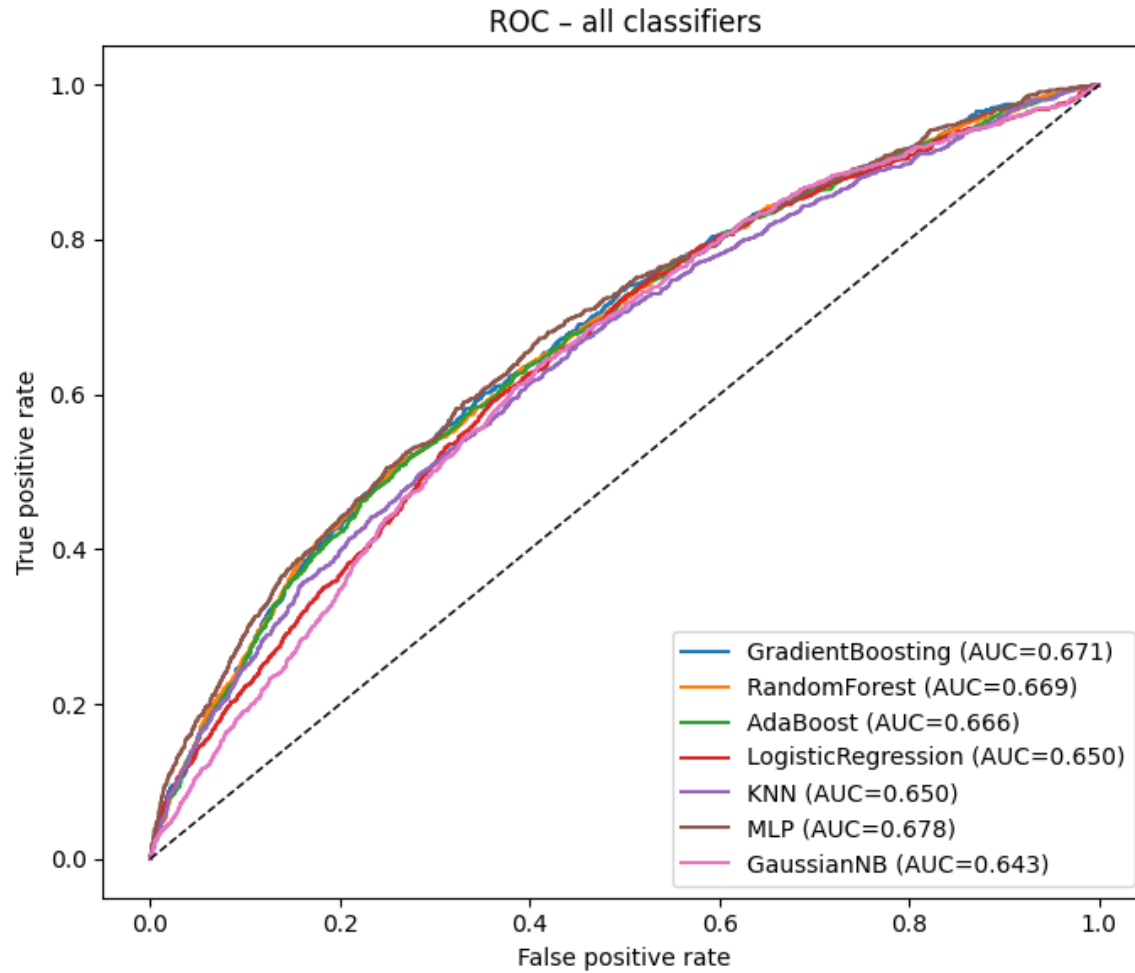
TMVA overtraining check



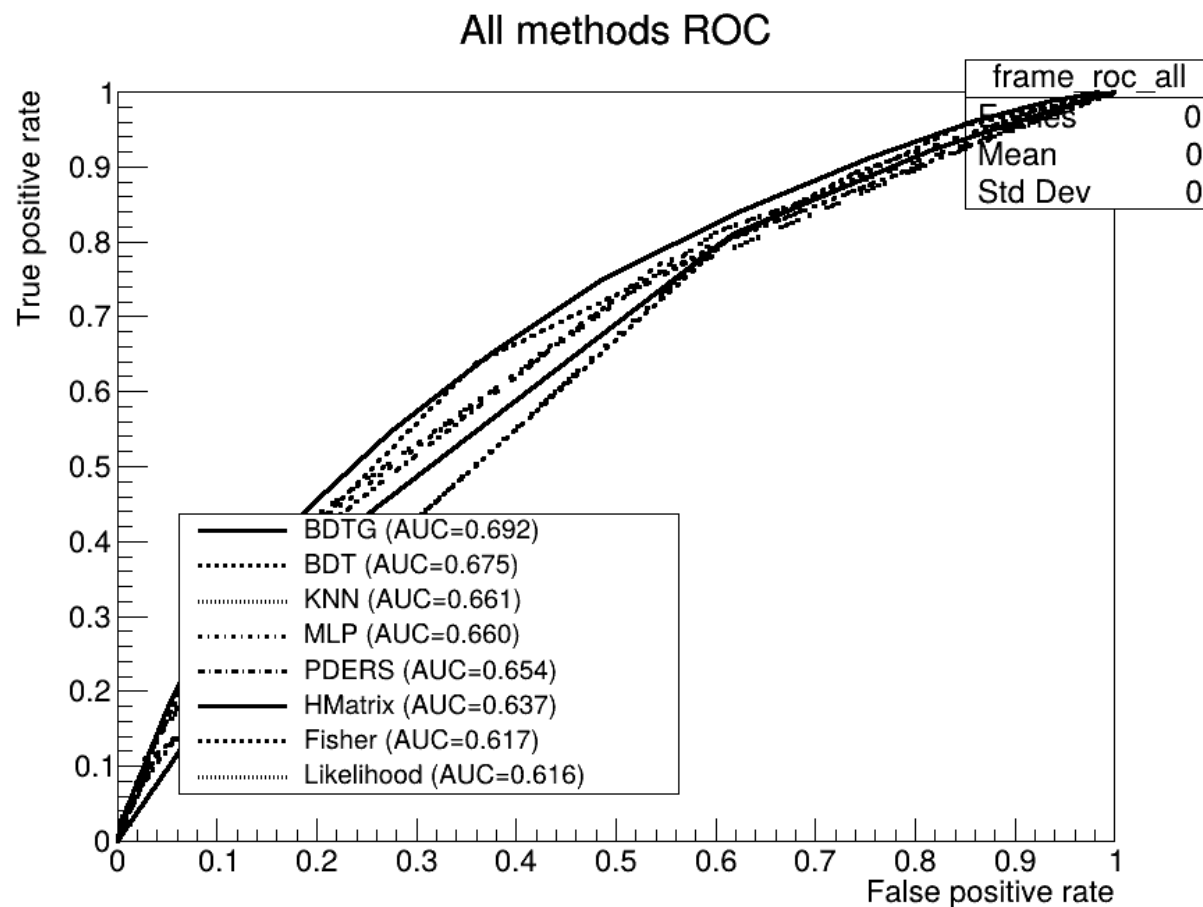
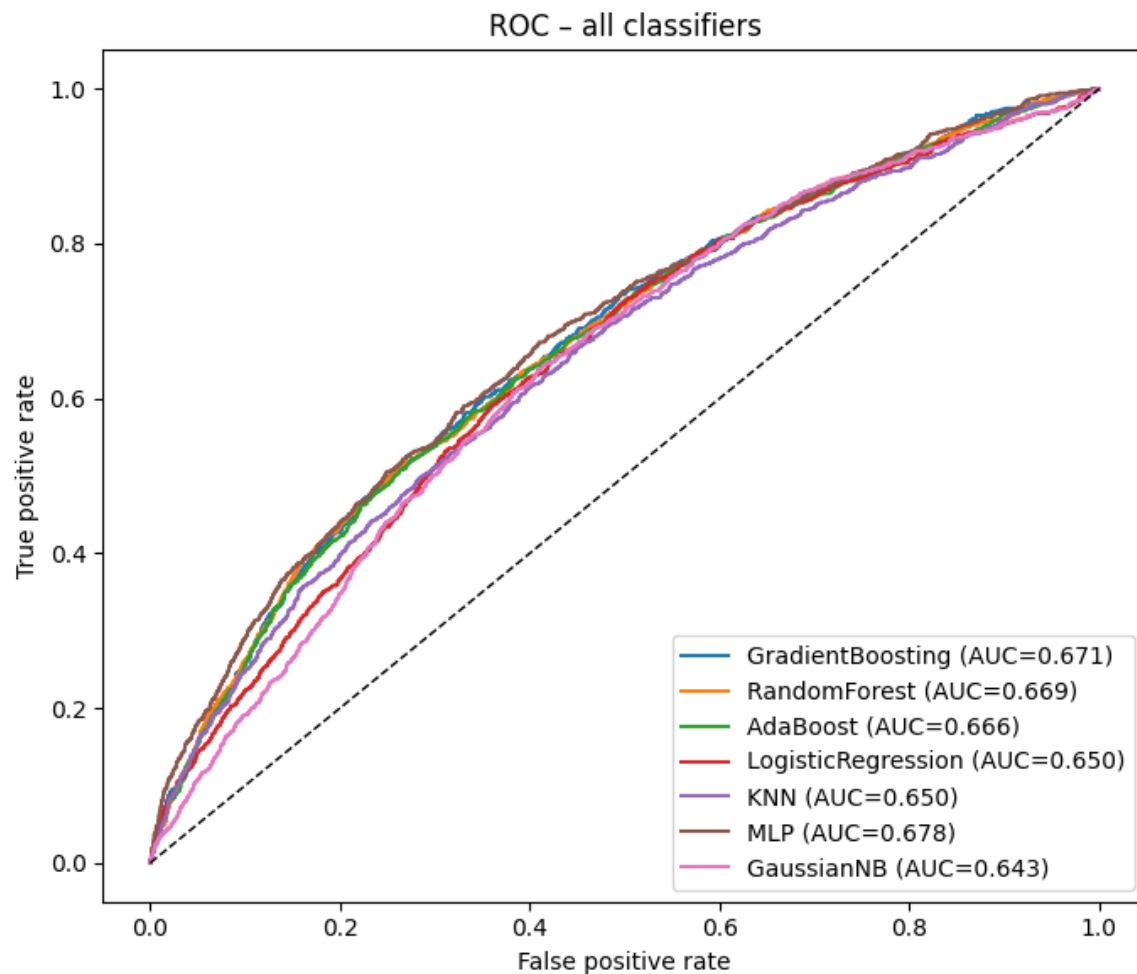
ROC of BDT from R and TMVA



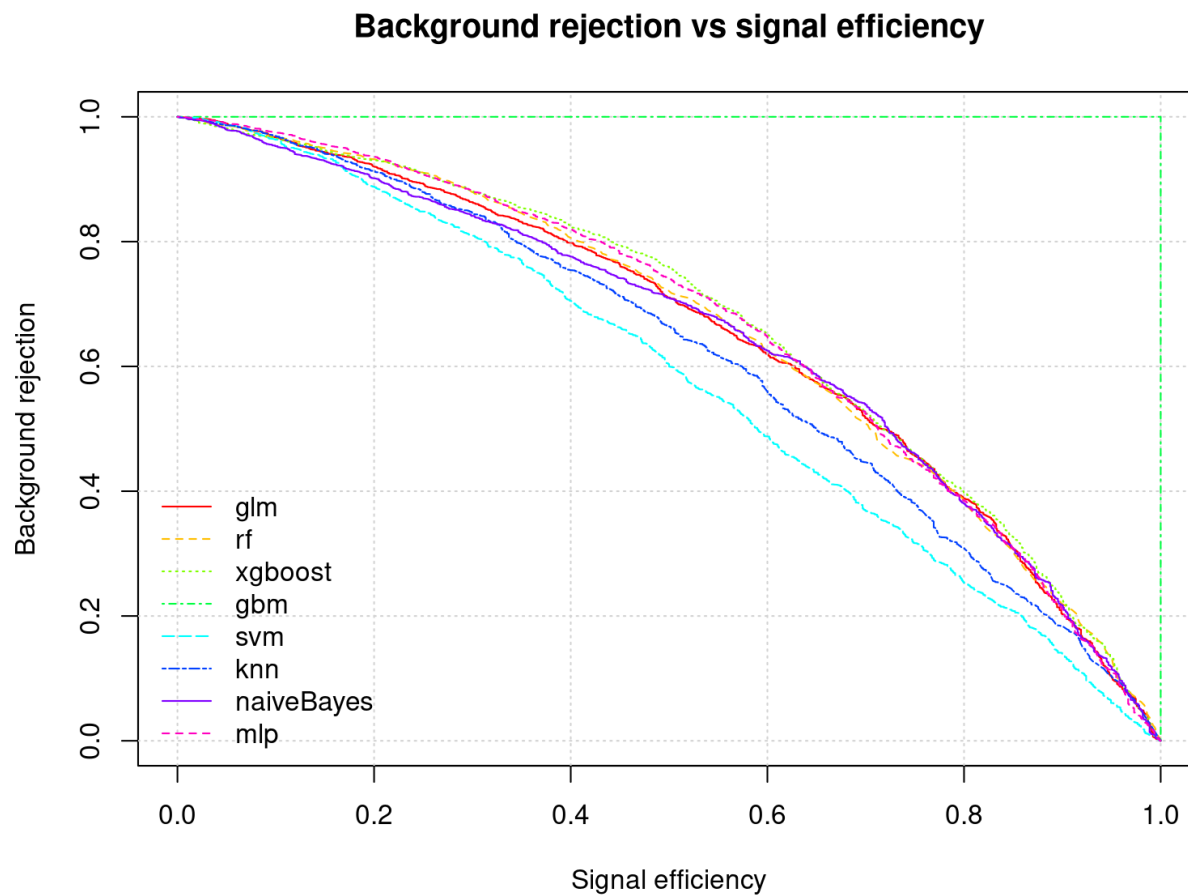
All classifiers ROC from sklearn and TMVA

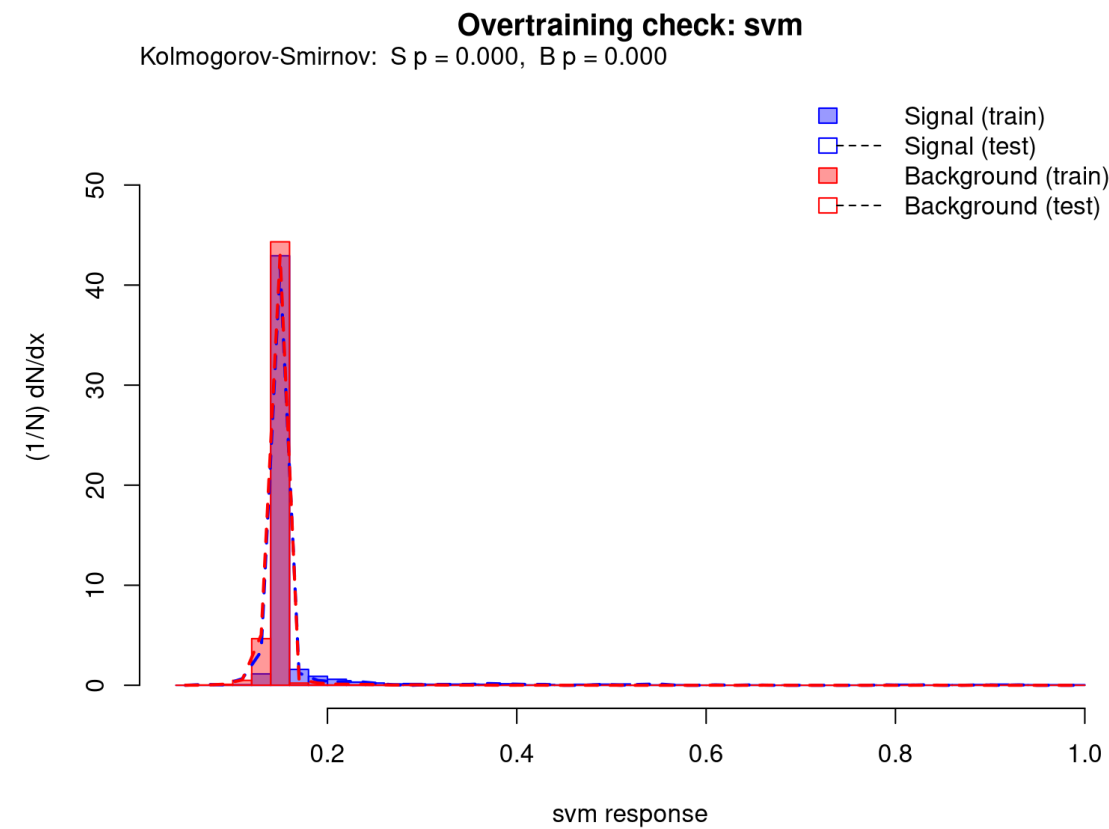
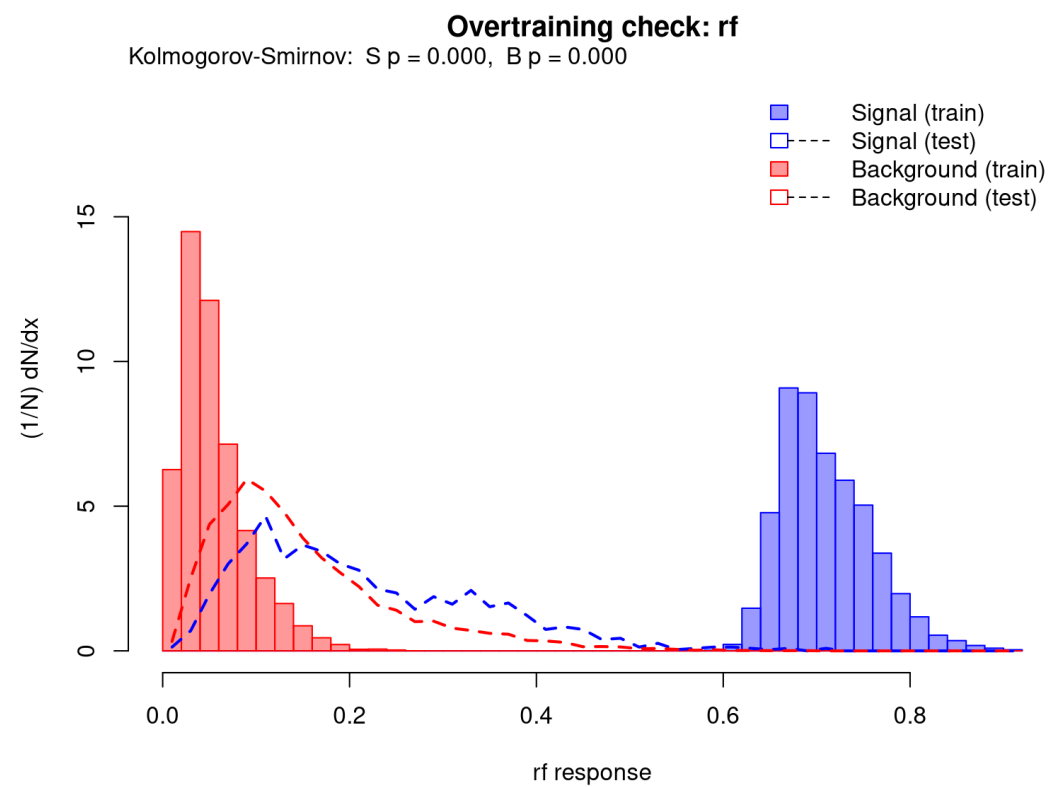


All classifiers ROC compared from R and TMVA environments



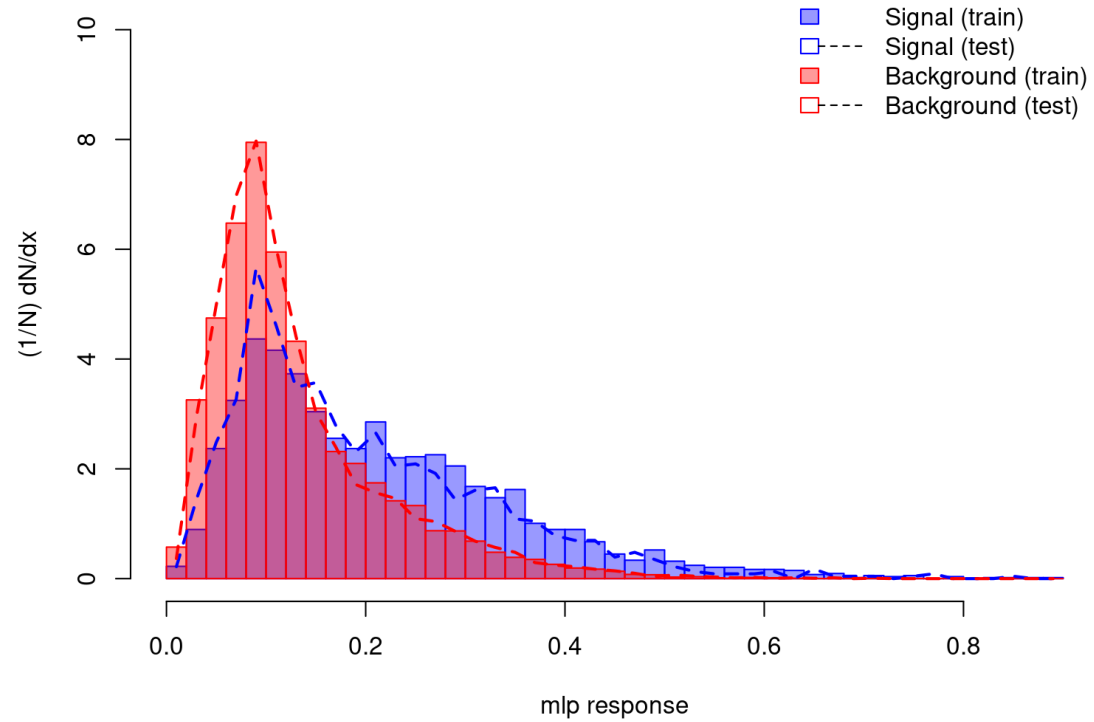
R based results





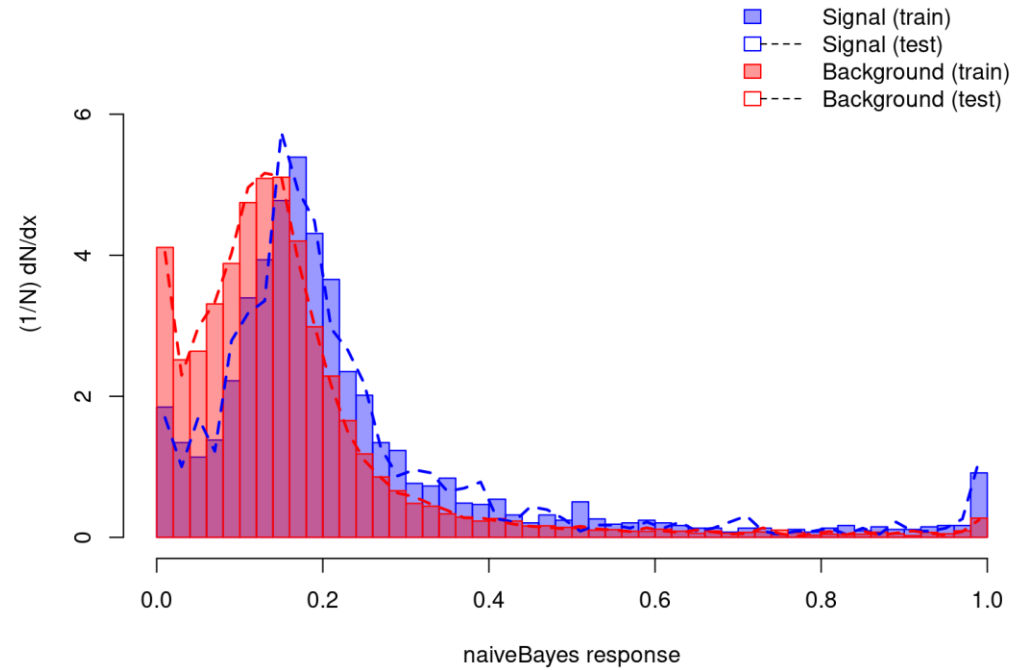
Overtraining check: mlp

Kolmogorov-Smirnov: S p = 0.005, B p = 0.518



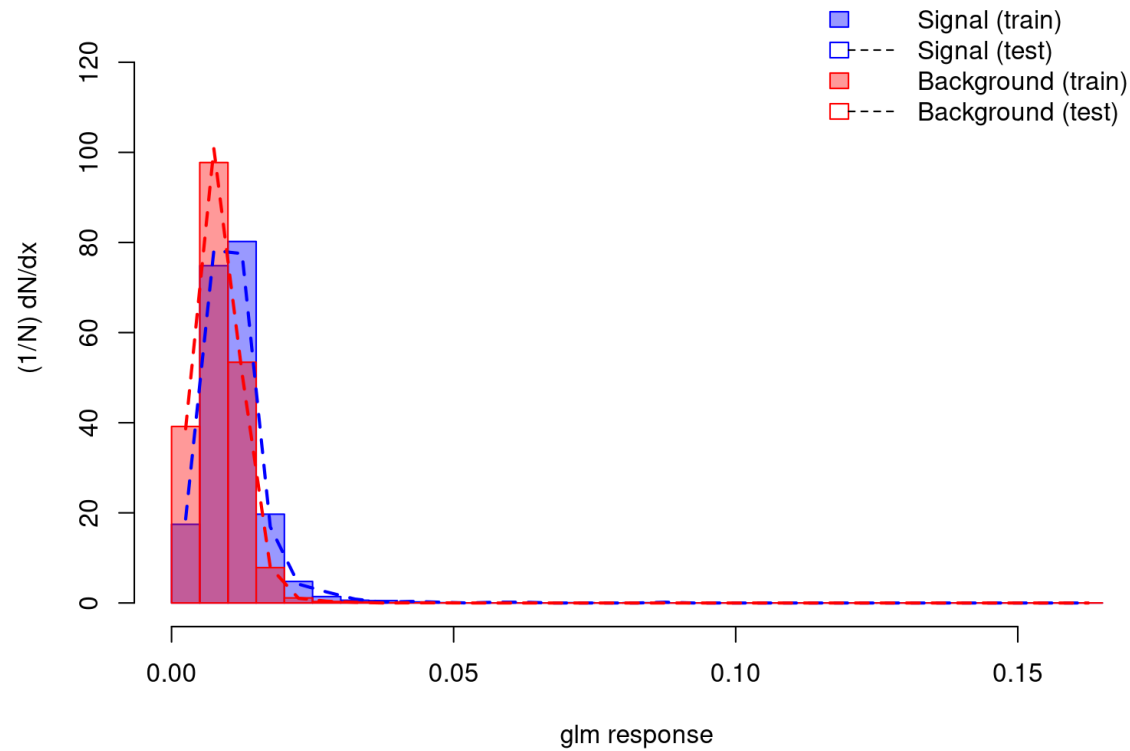
Overtraining check: naiveBayes

Kolmogorov-Smirnov: S p = 0.982, B p = 0.384



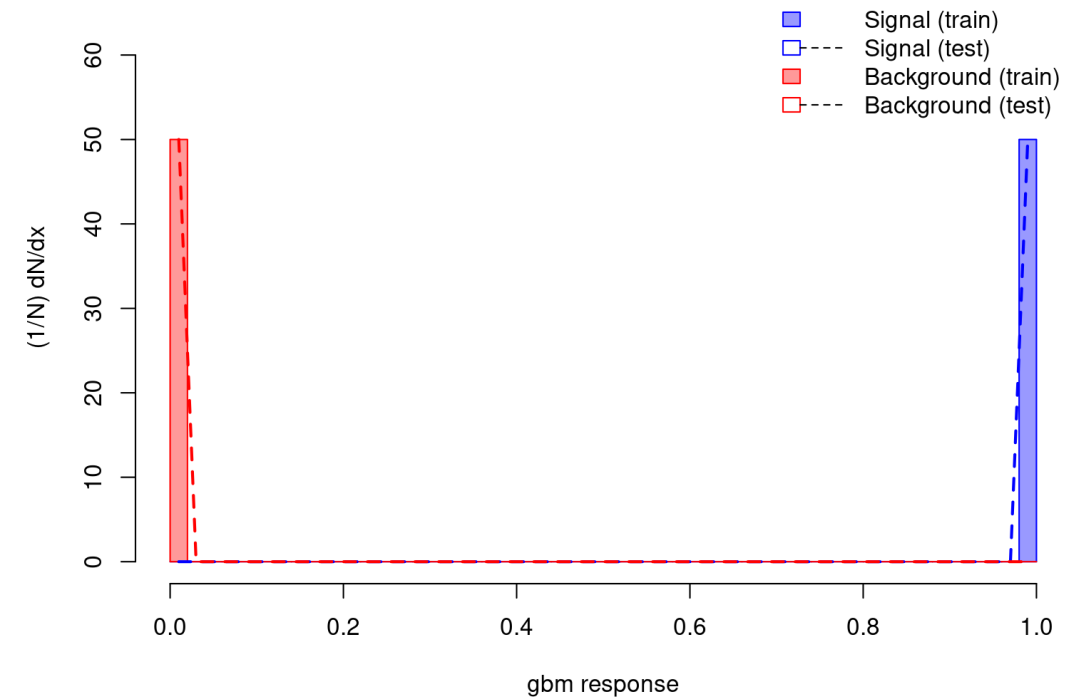
Overtraining check: glm

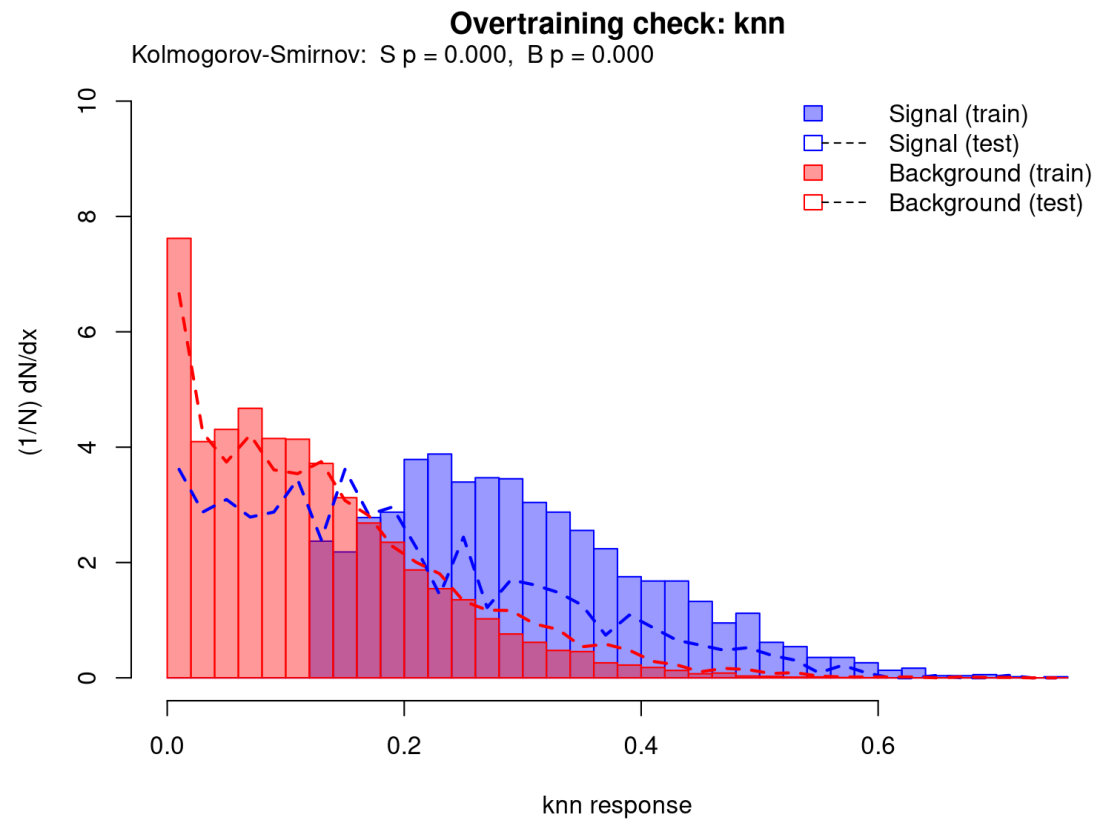
Kolmogorov-Smirnov: S p = 0.680, B p = 0.112



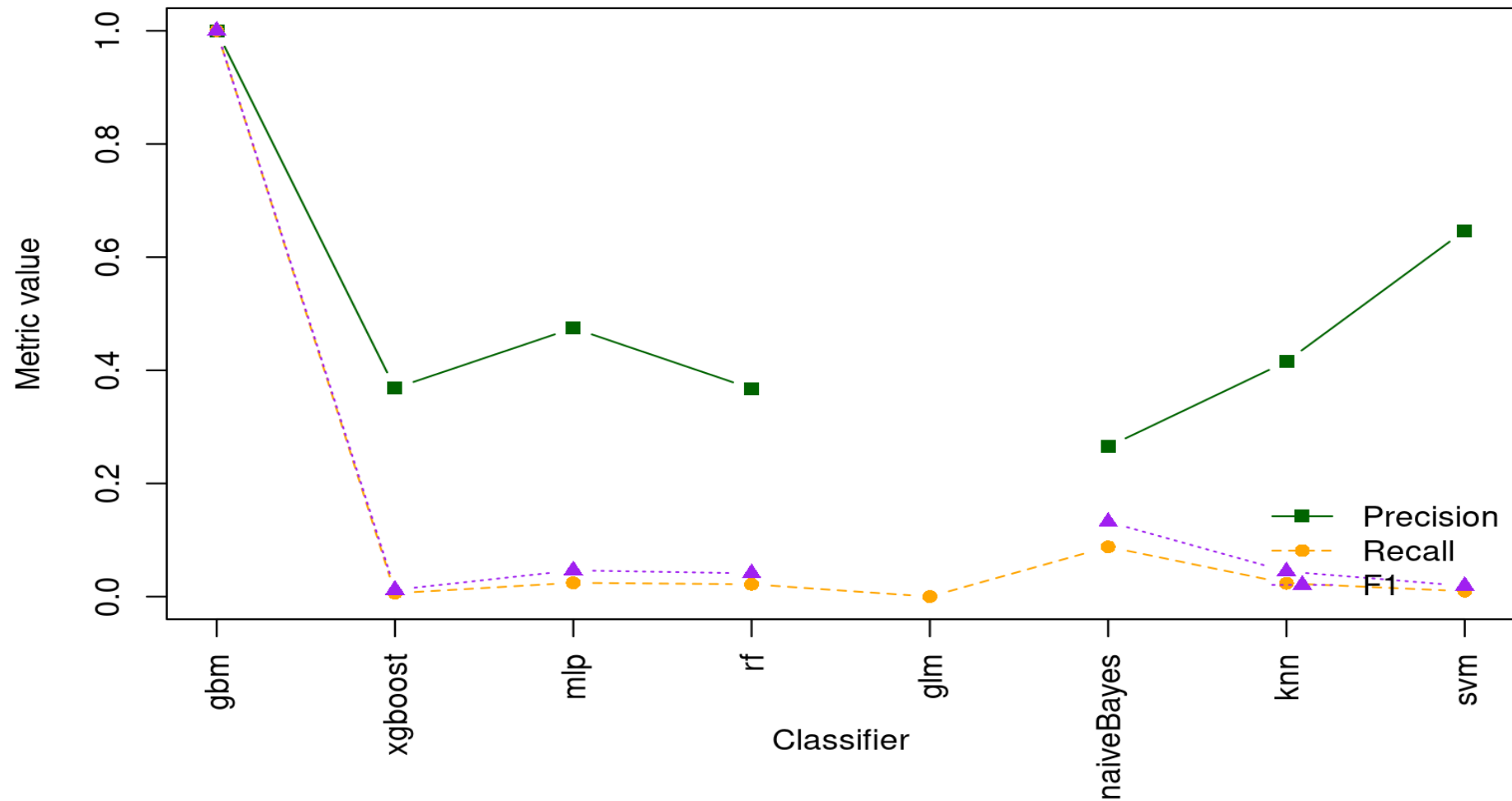
Overtraining check: gbm

Kolmogorov-Smirnov: S p = 1.000, B p = 1.000

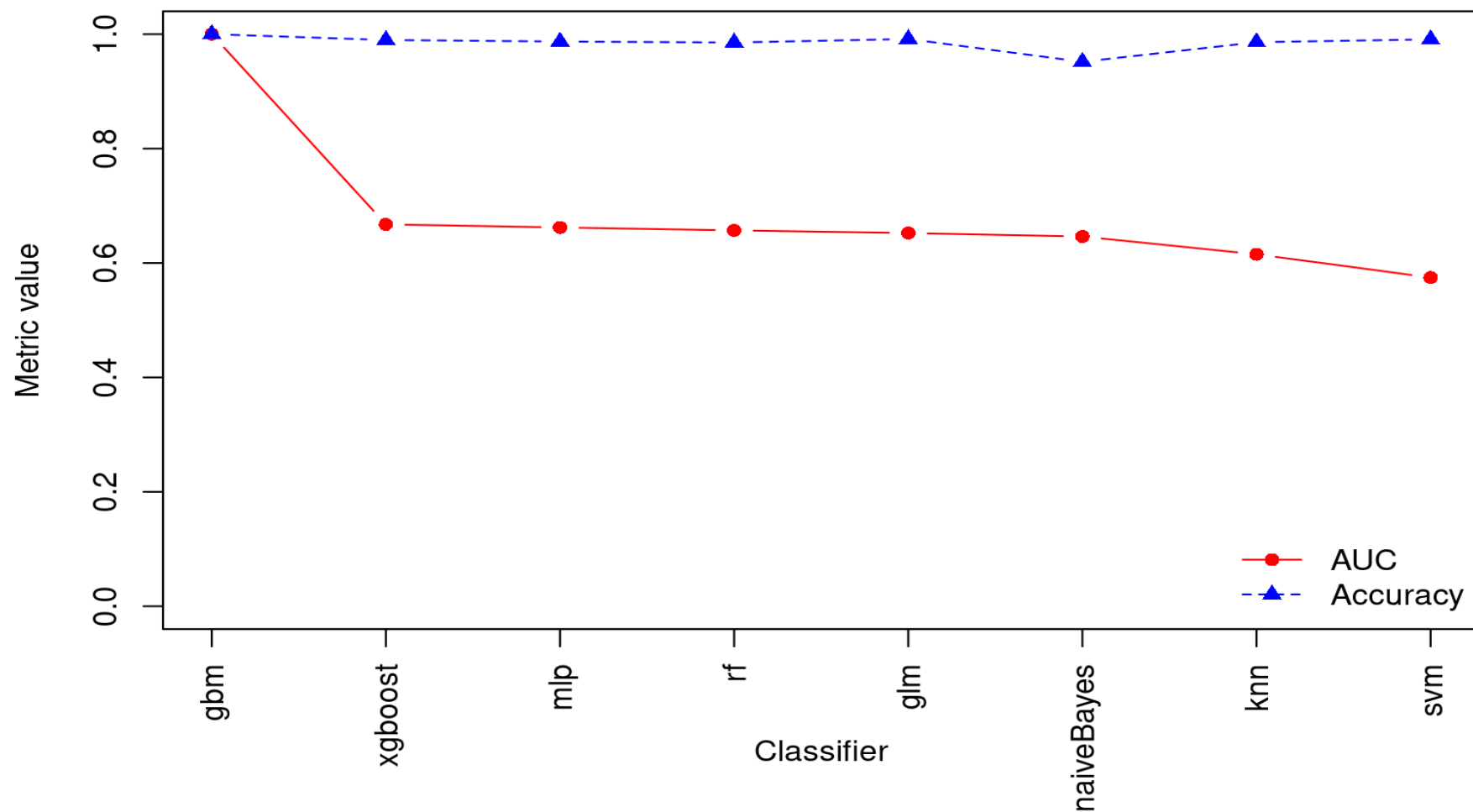




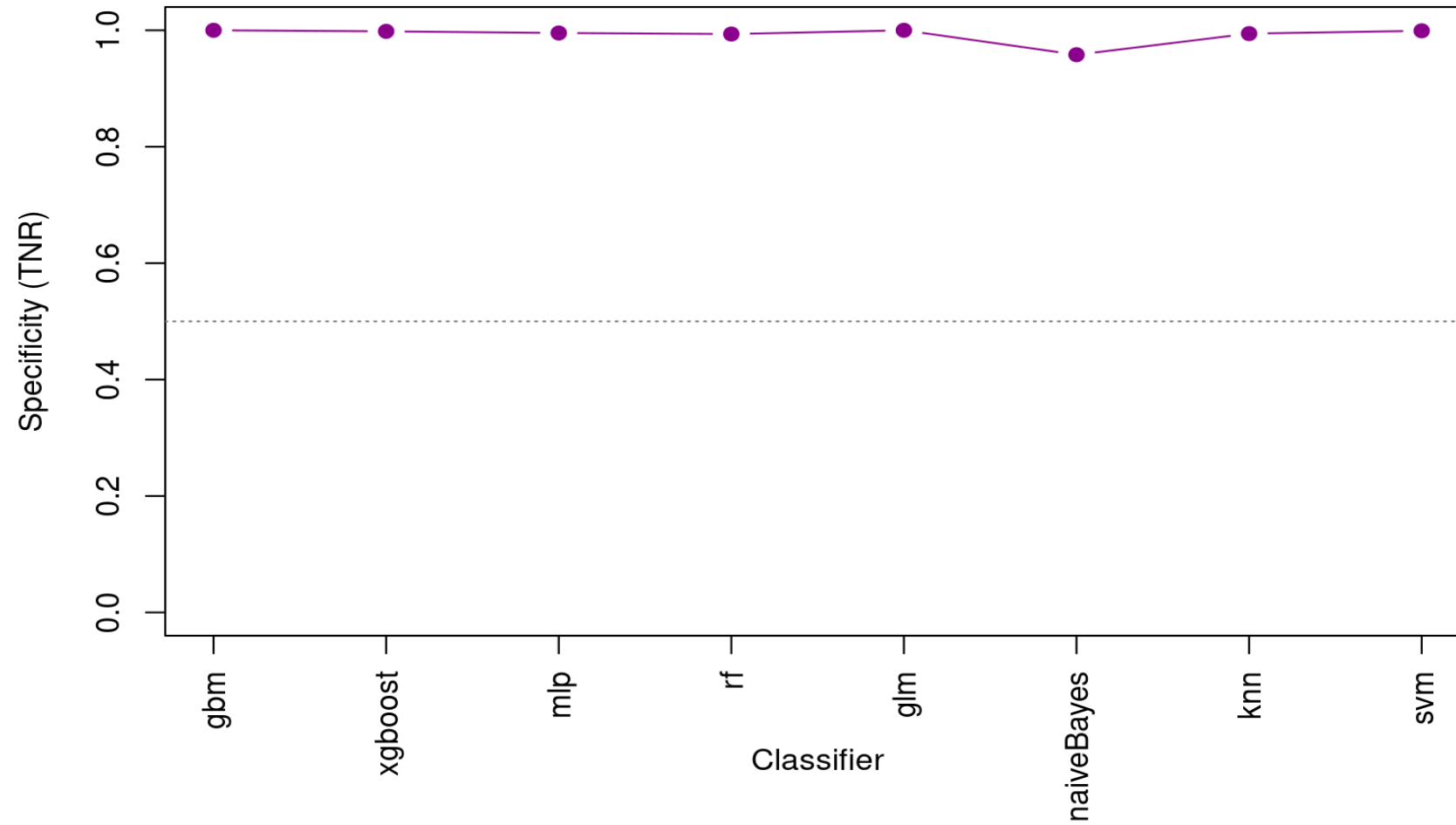
Precision, recall, F1 (test sample)



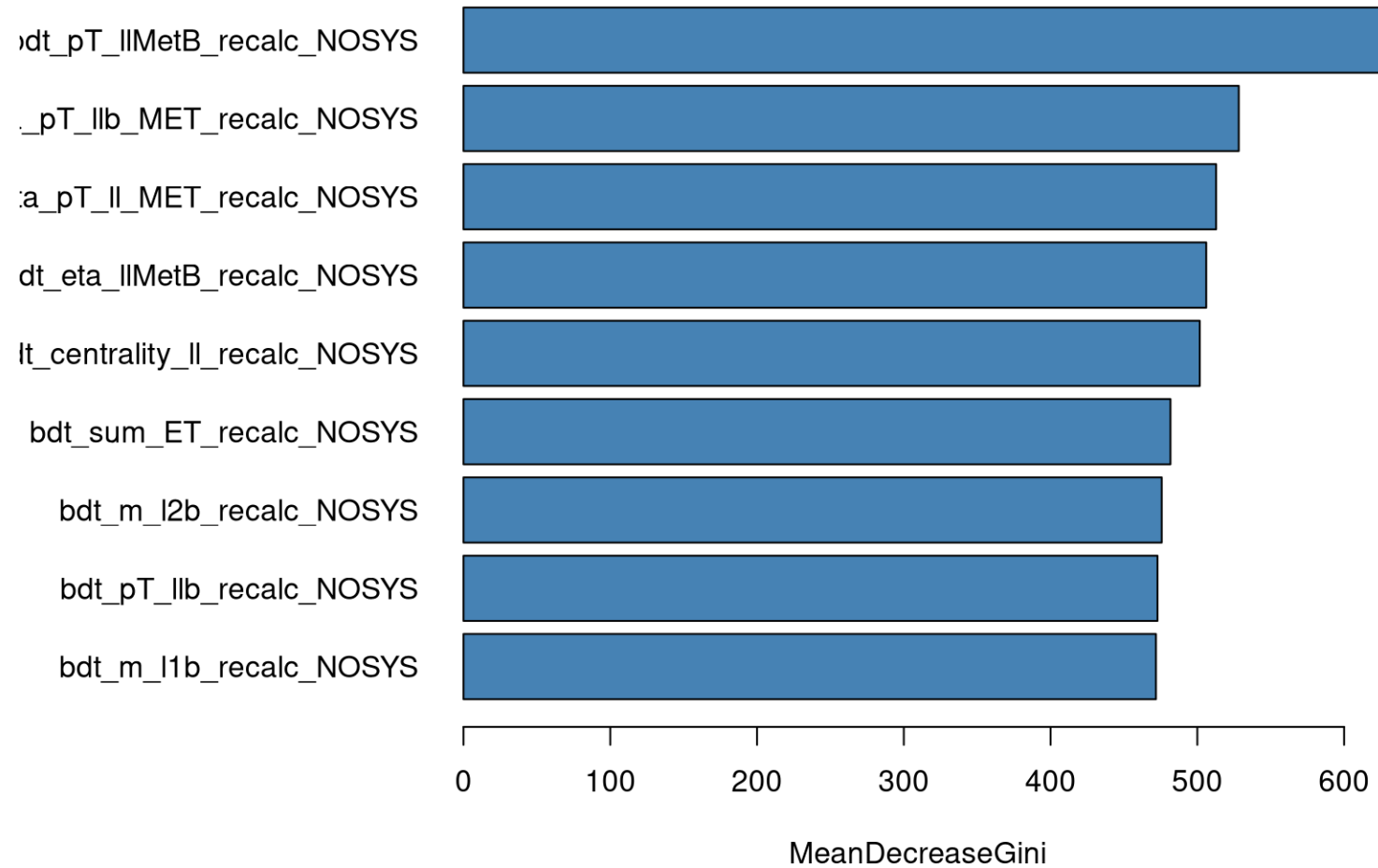
AUC and accuracy (test sample)



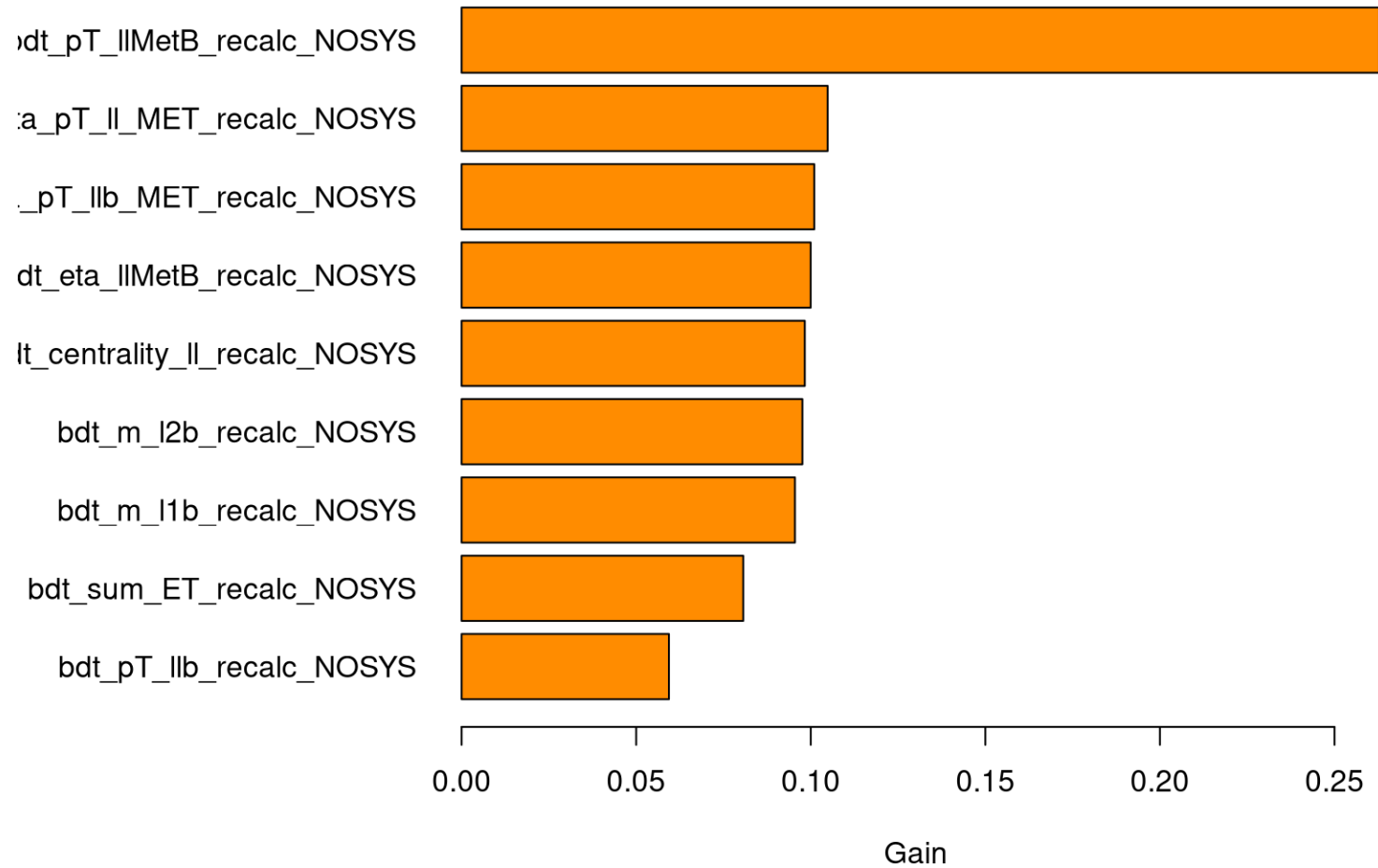
Specificity on test sample



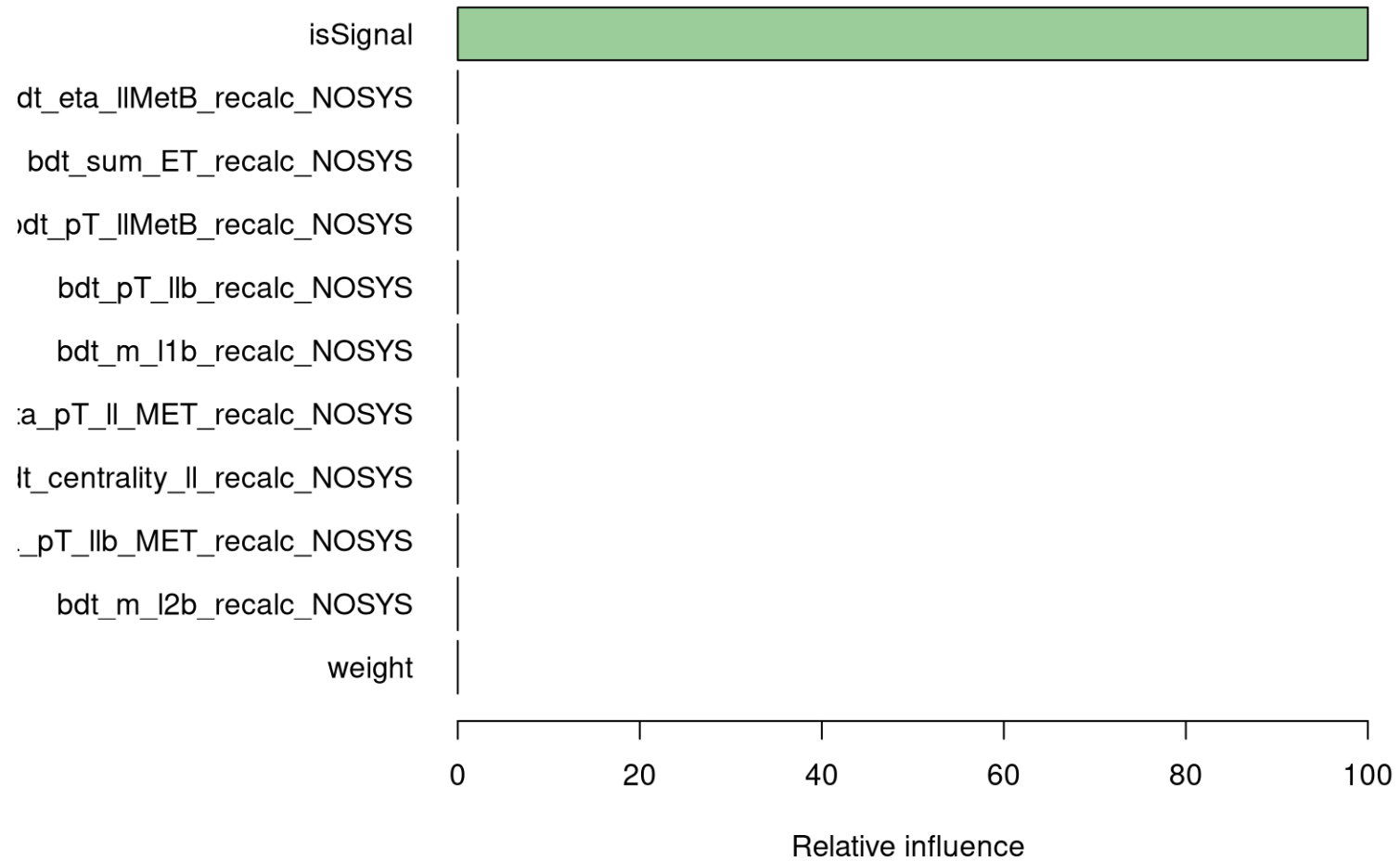
Random Forest variable importance



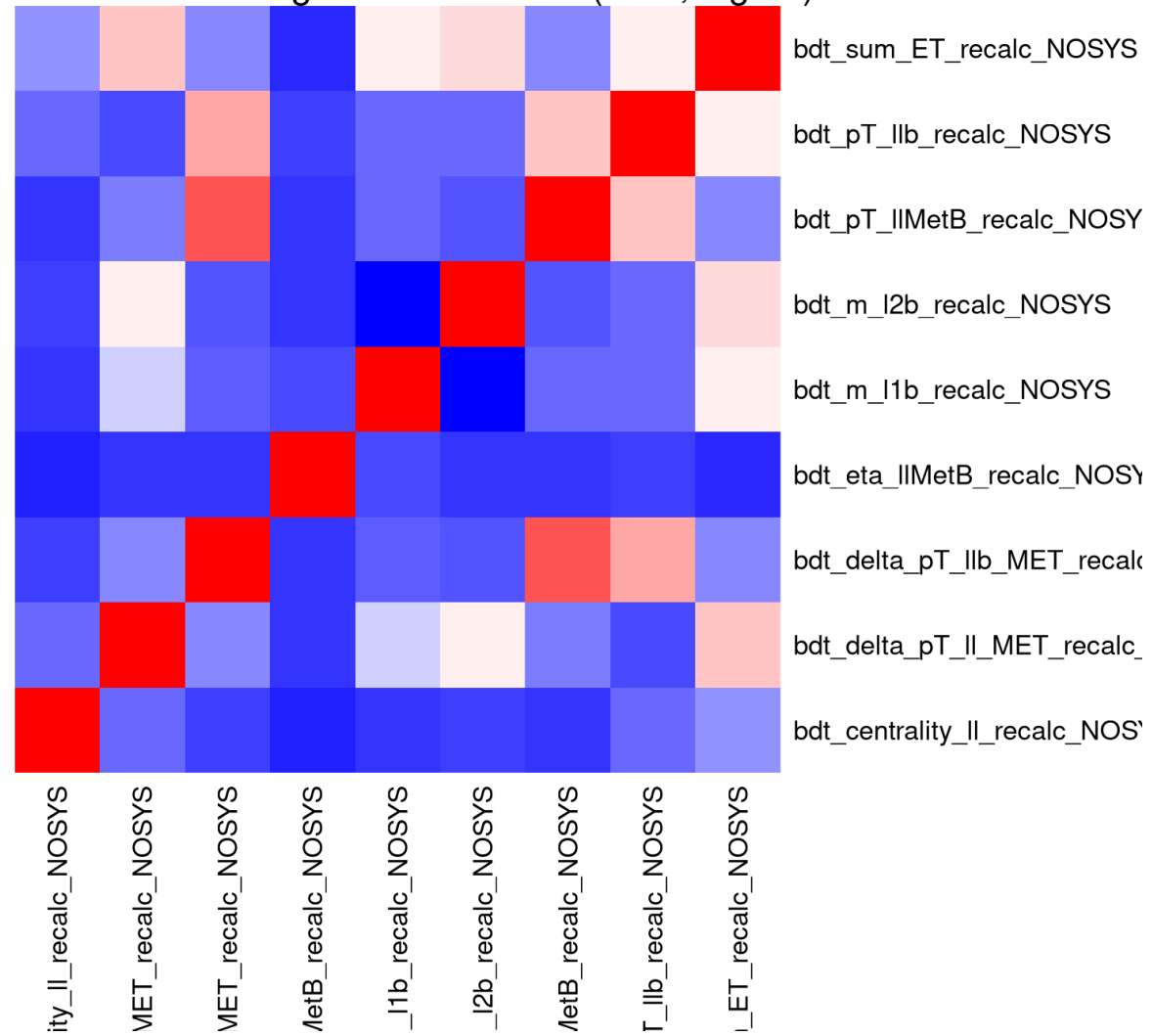
XGBoost variable importance (Gain)

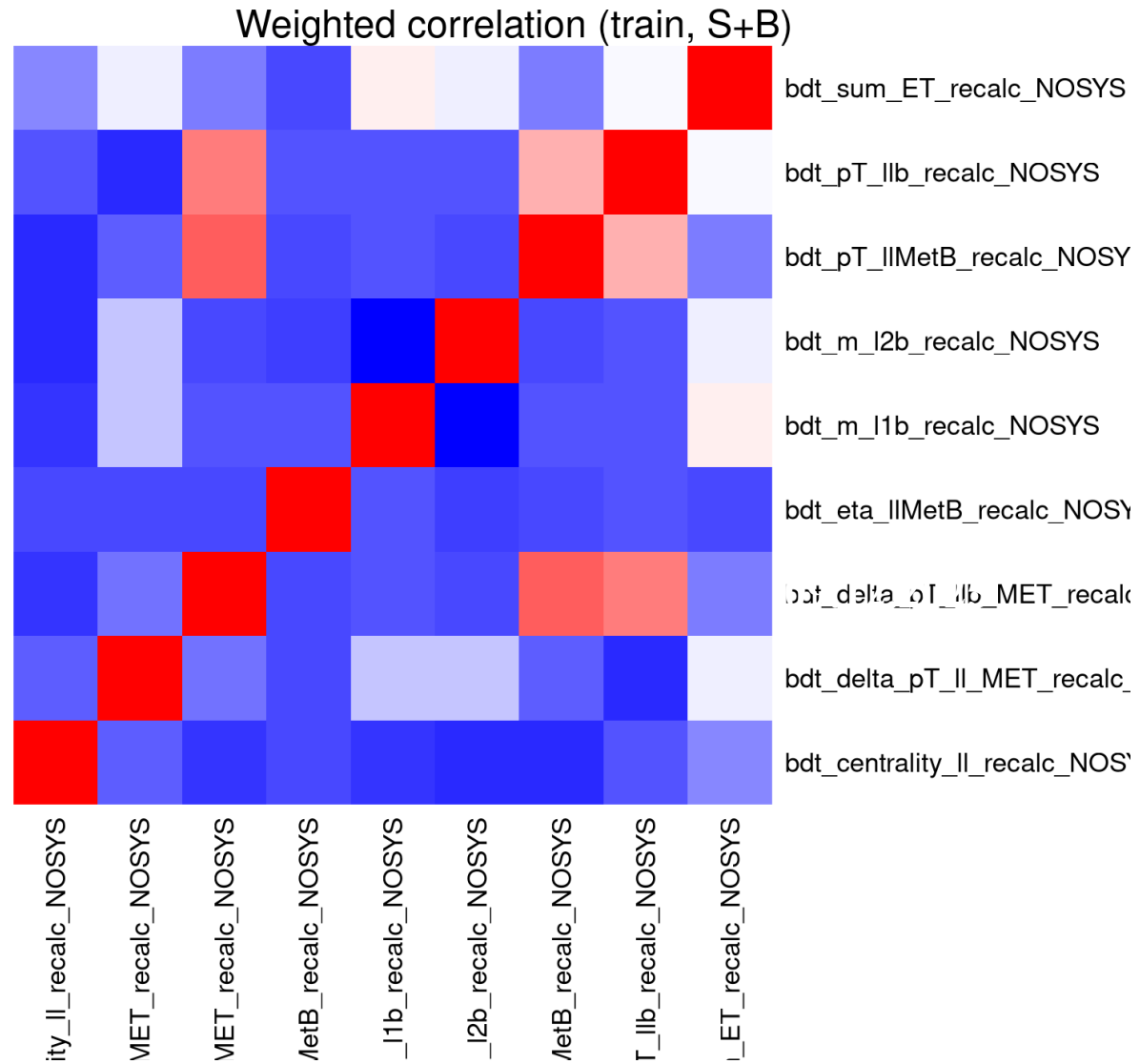


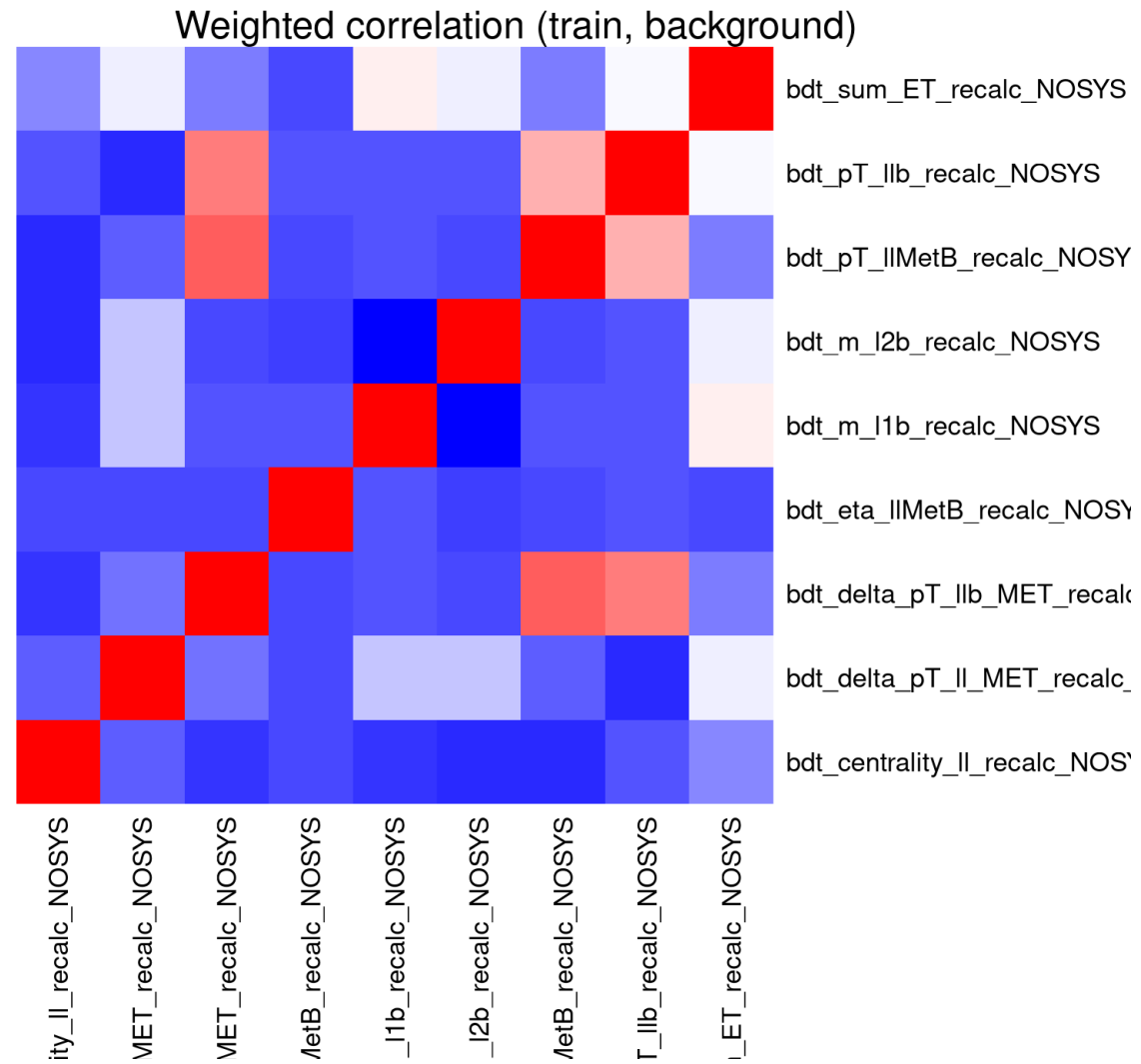
GBM variable importance



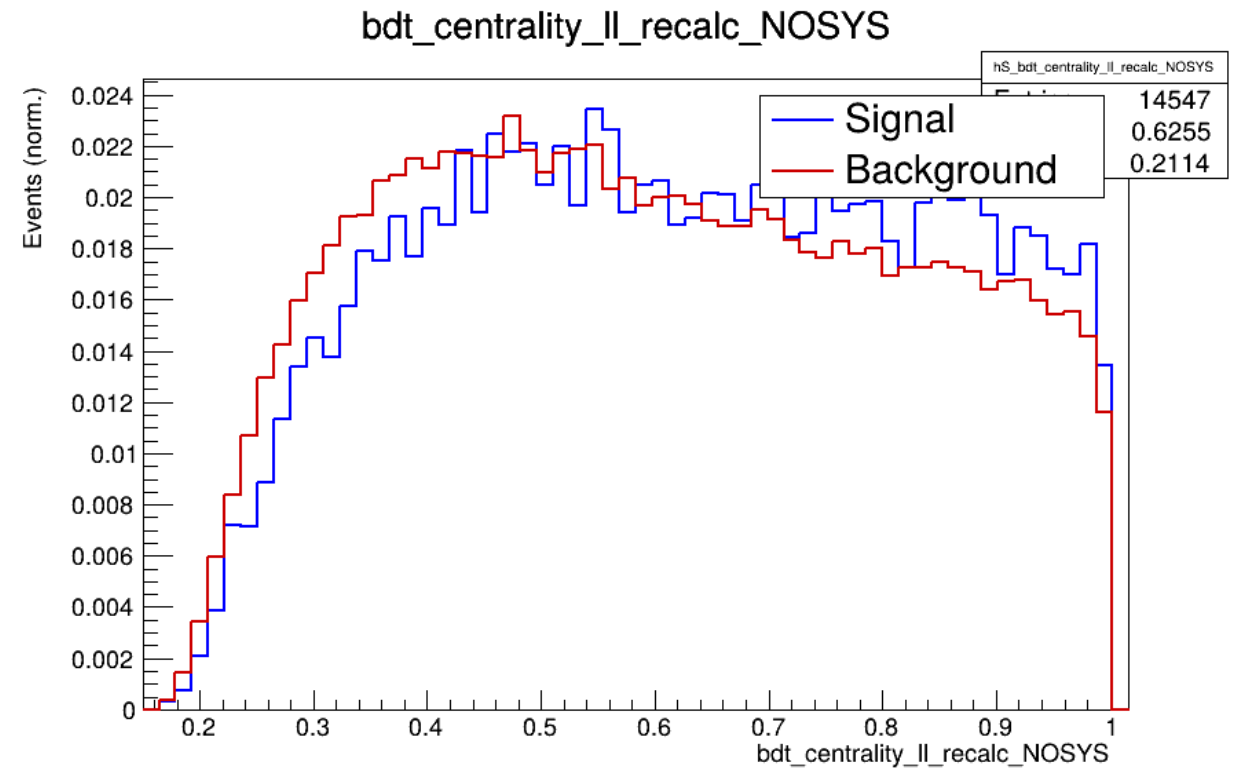
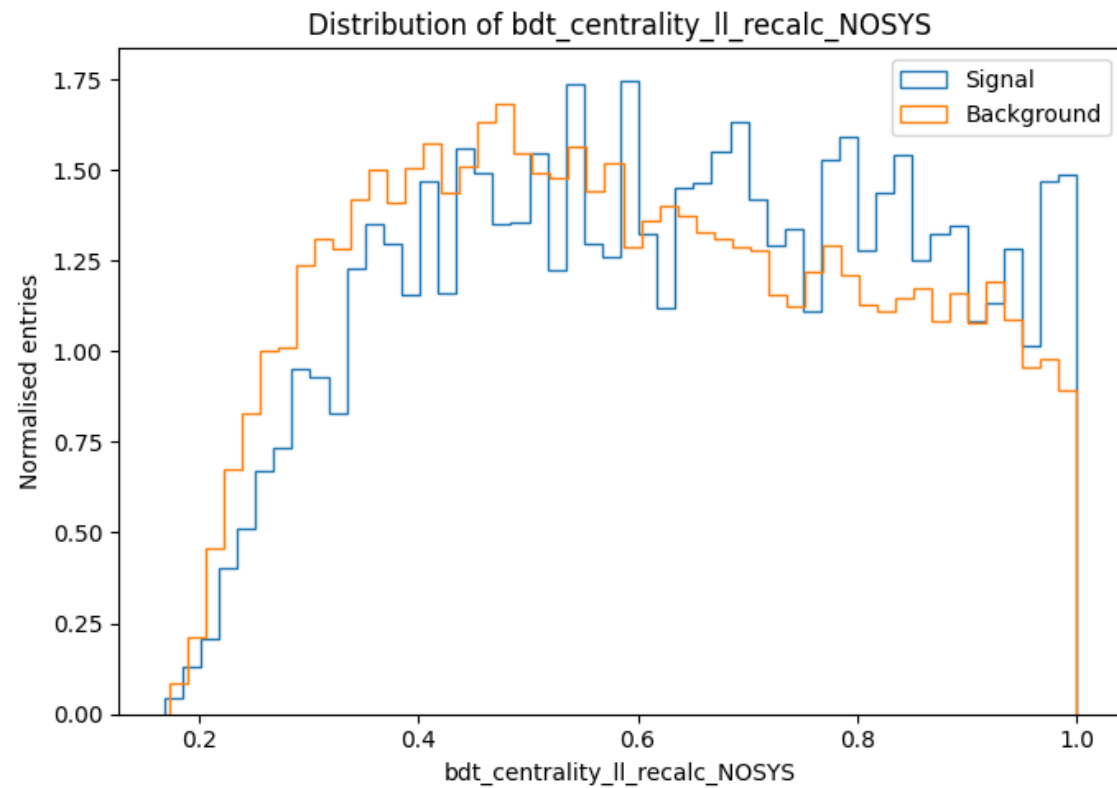
Weighted correlation (train, signal)

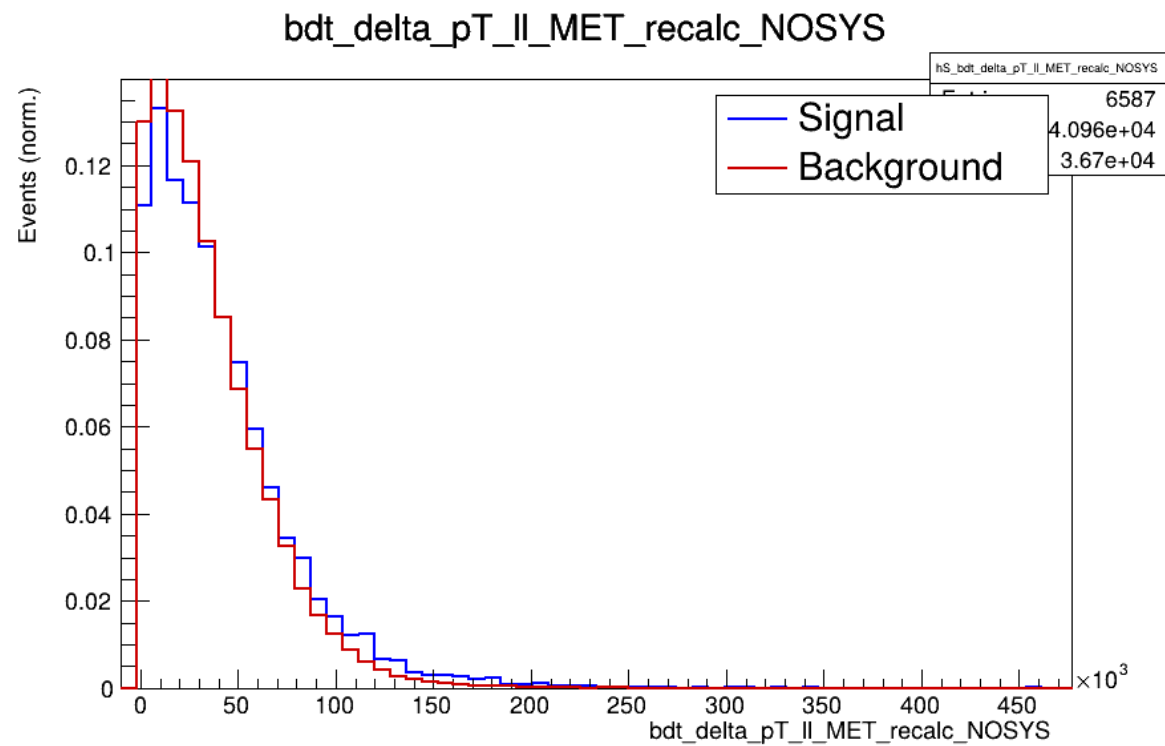
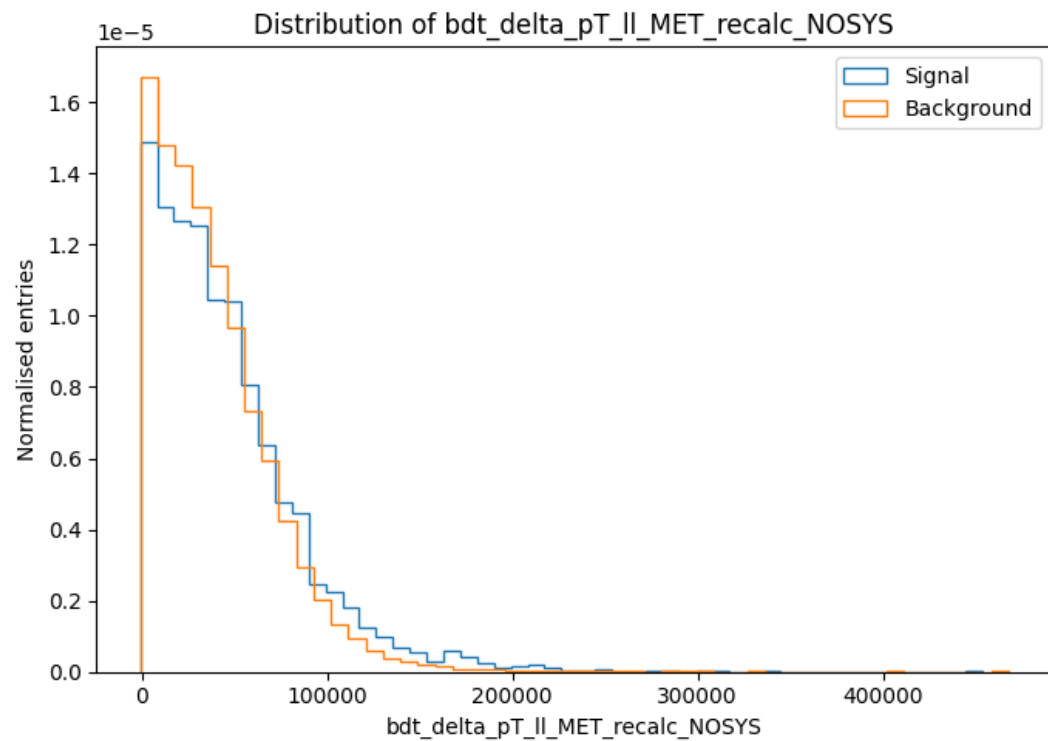


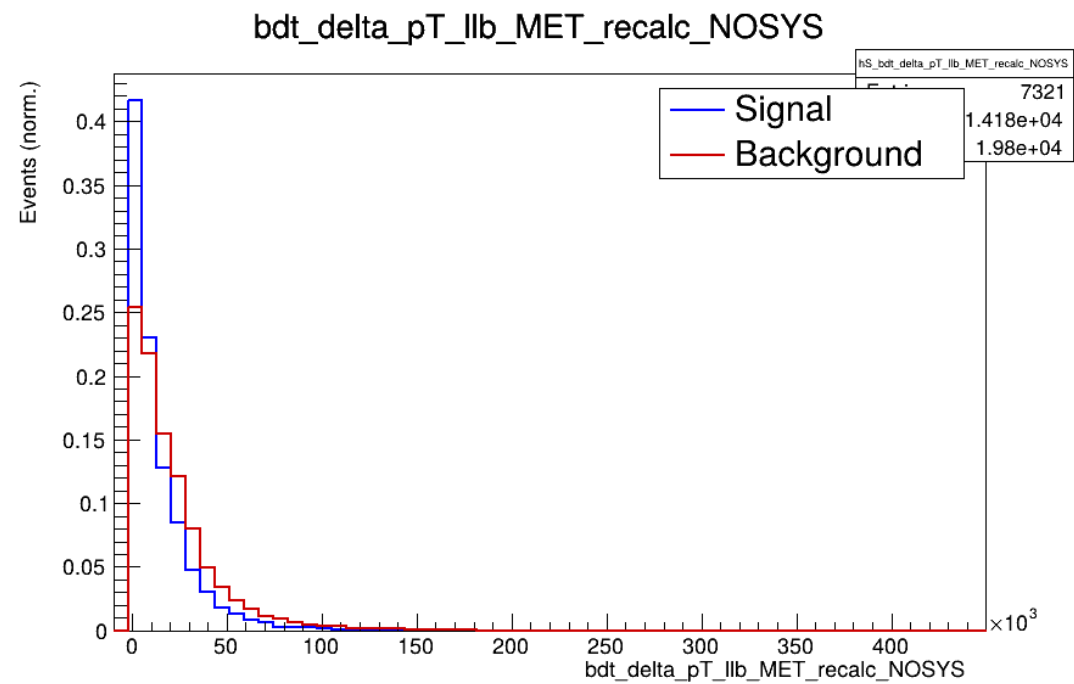
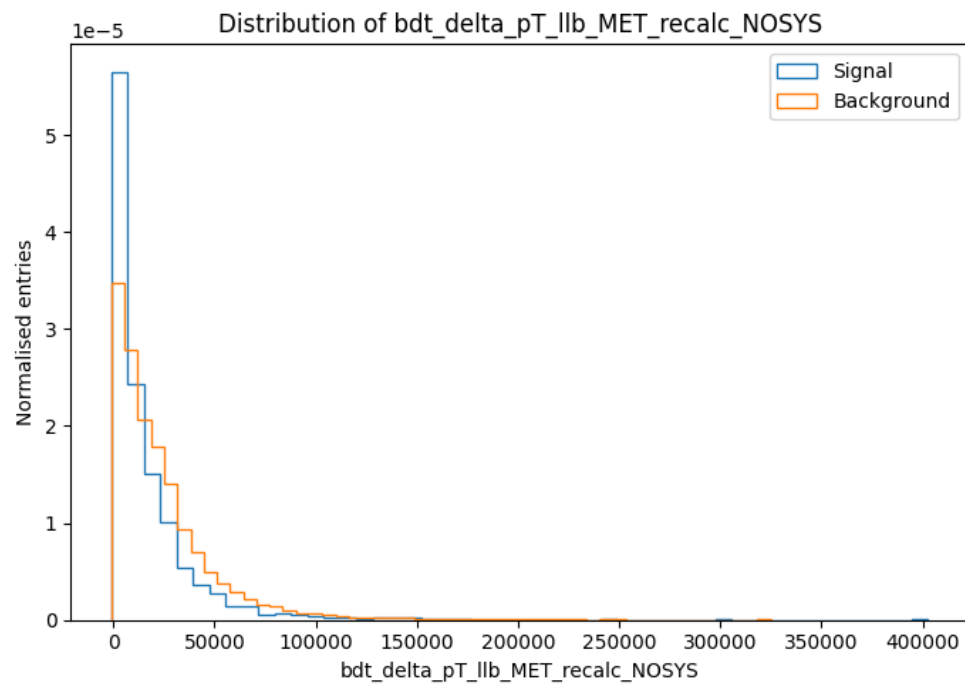


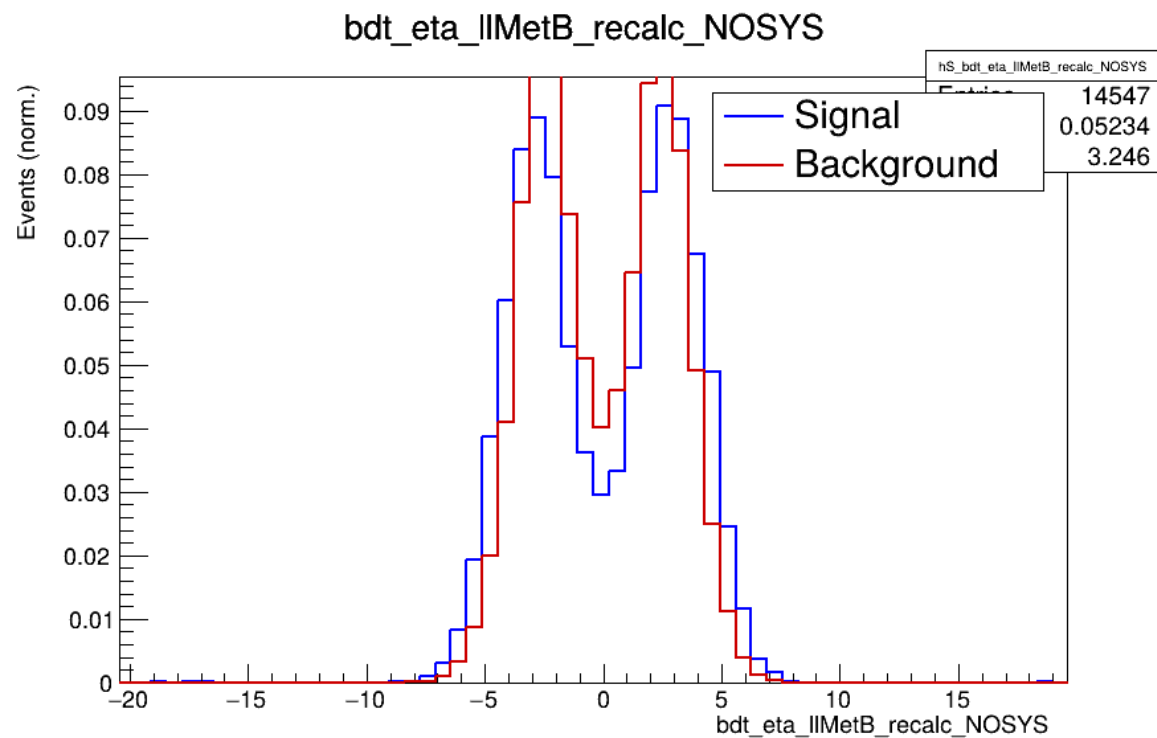
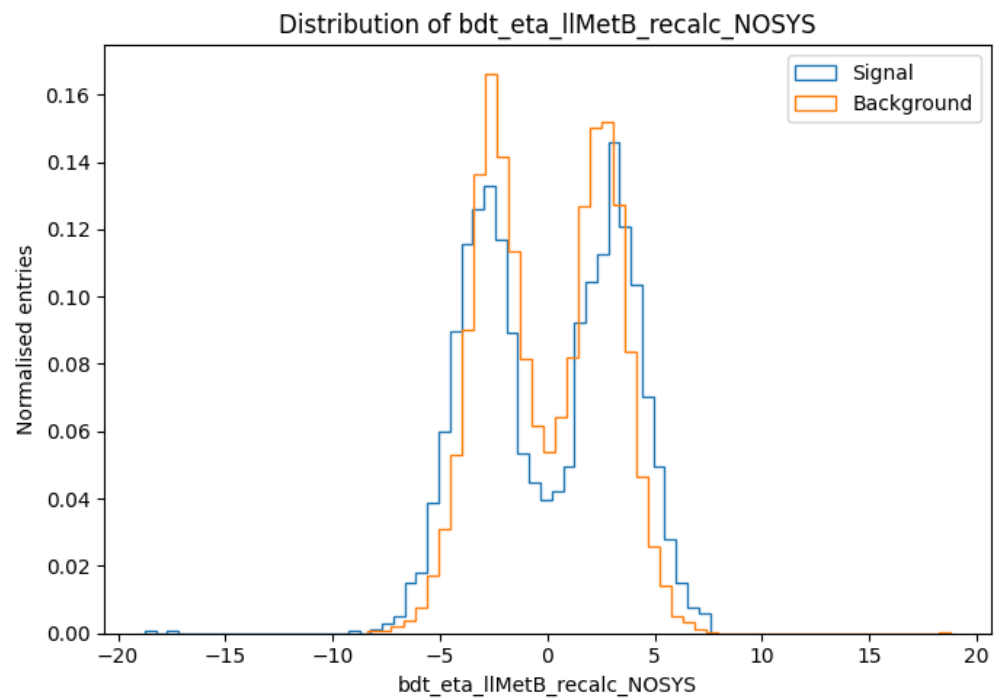


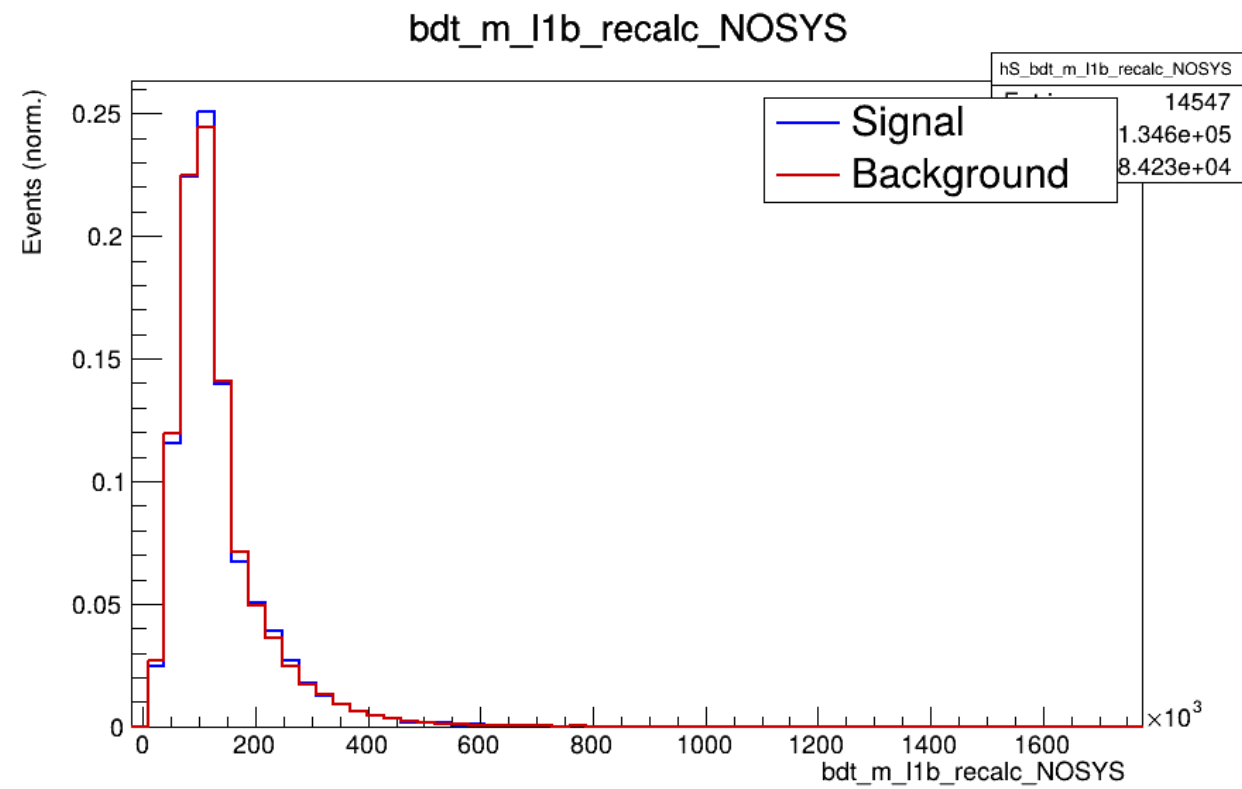
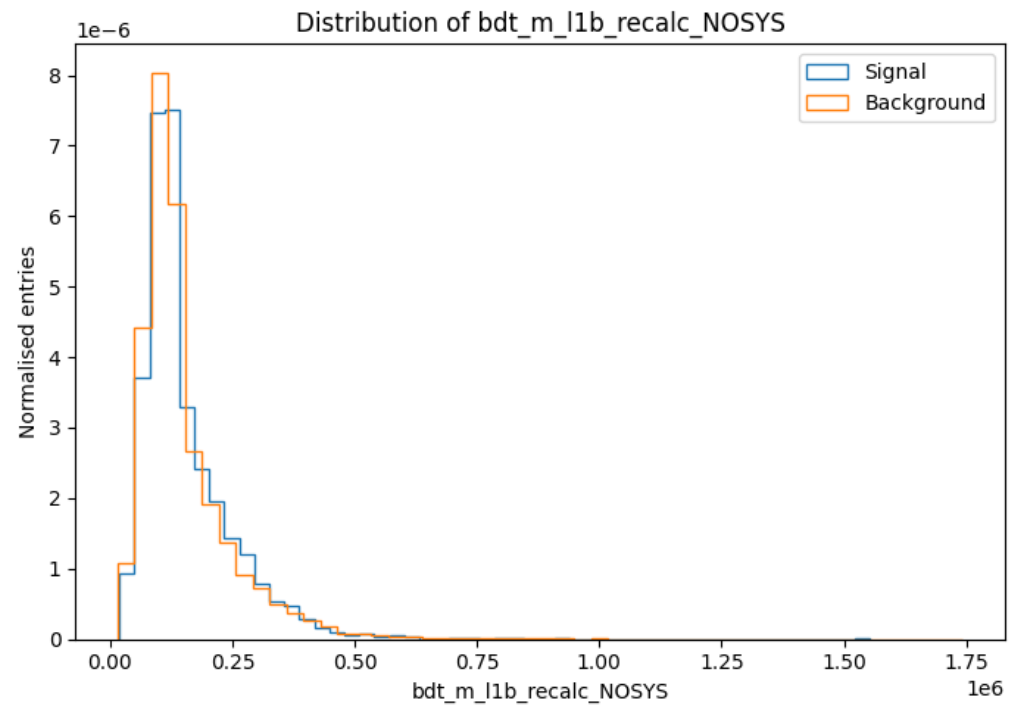
Variable separations comparison of plots in sklearn and TMVA environment

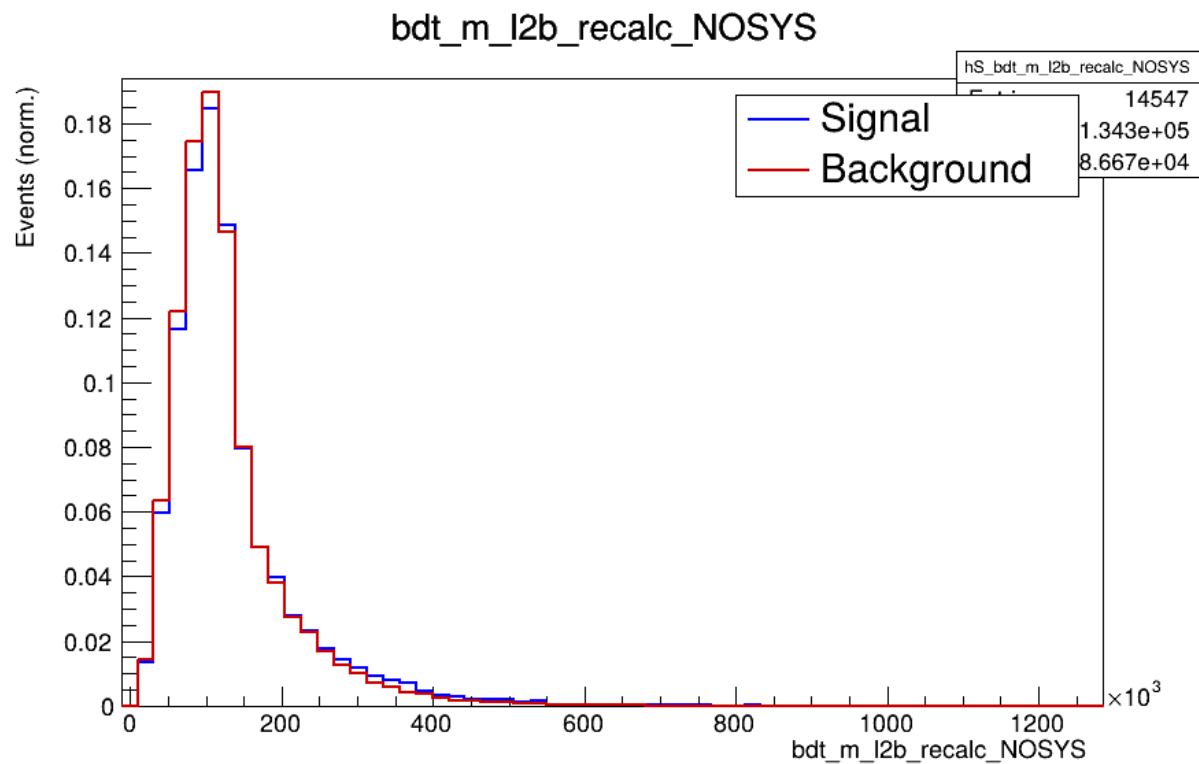
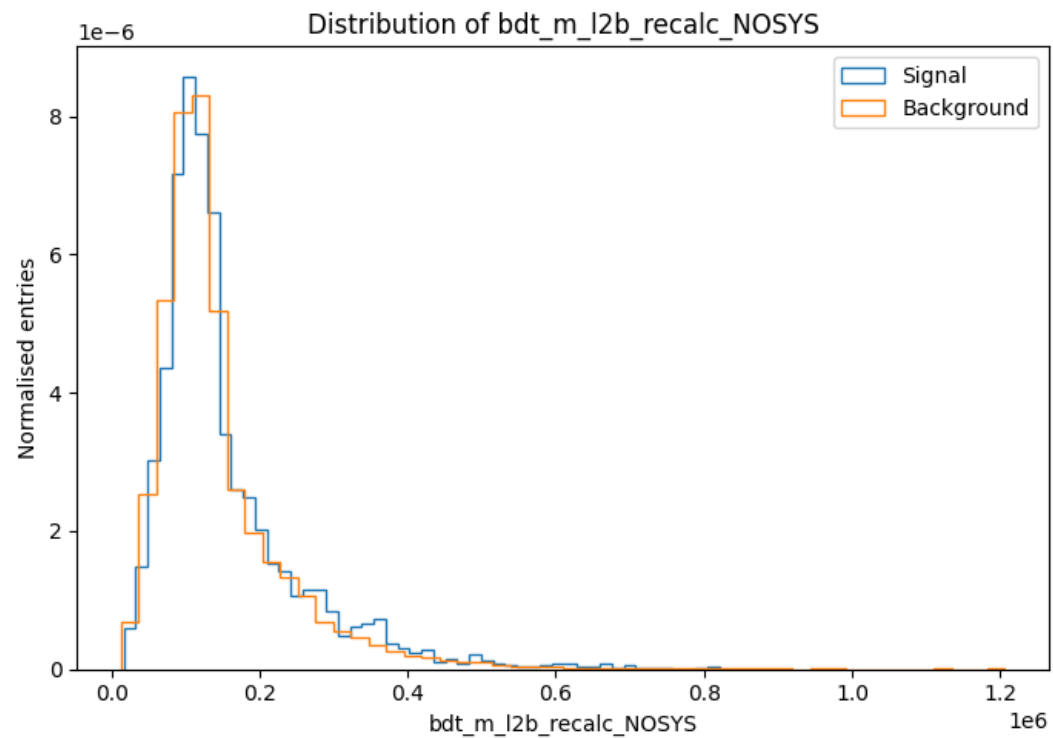


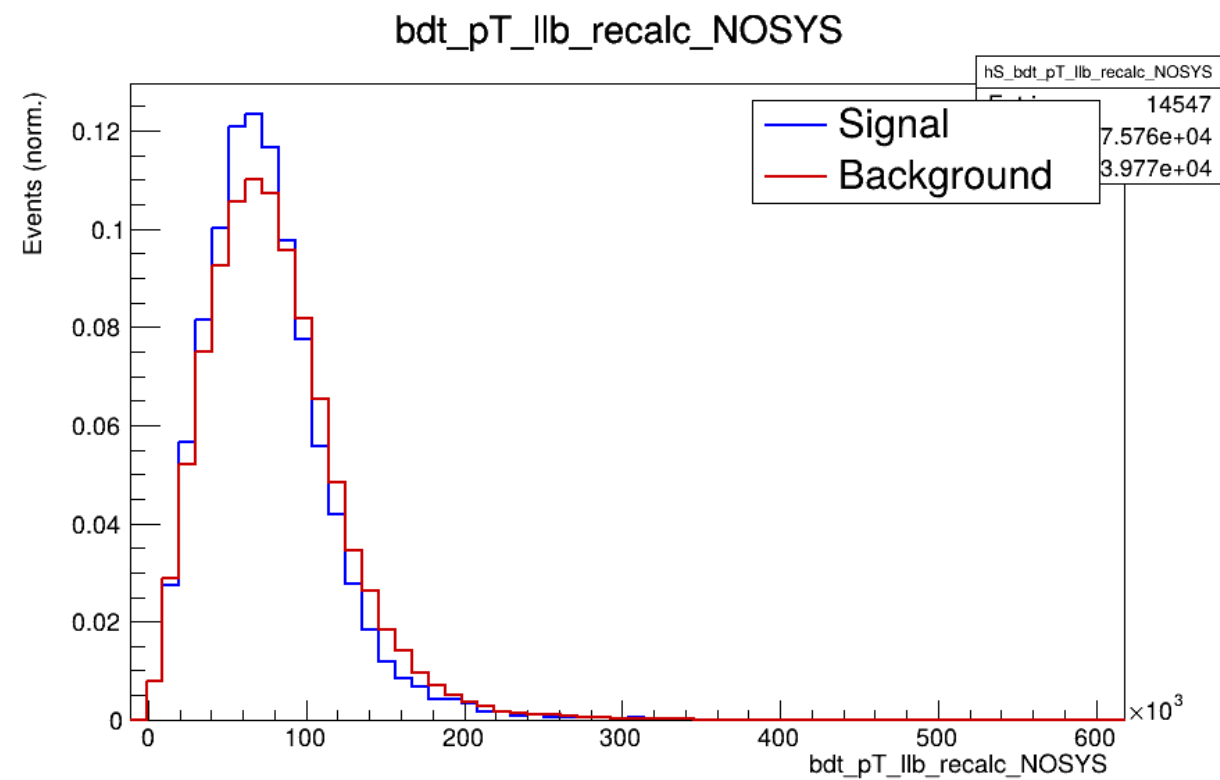
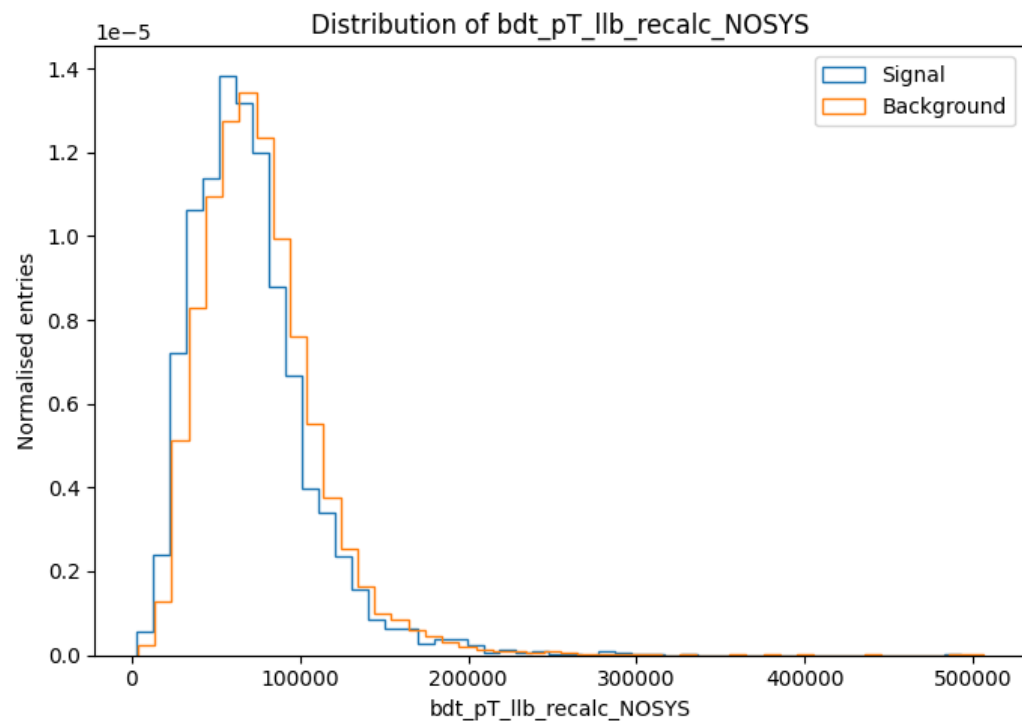


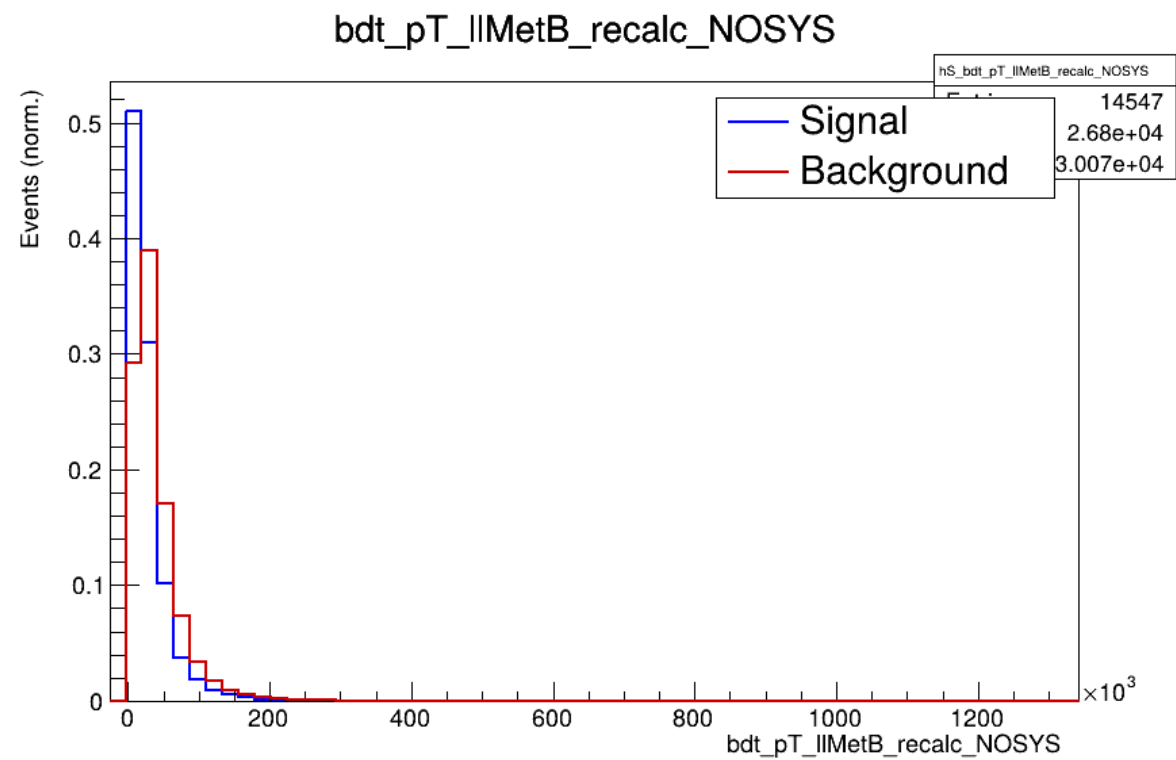
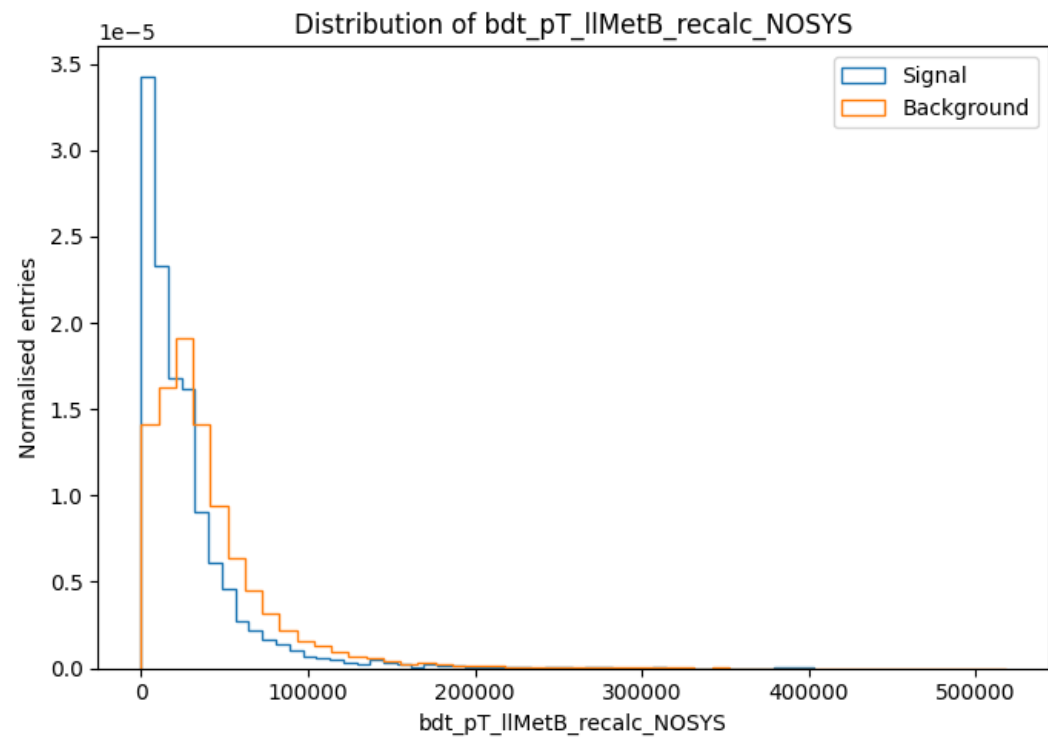






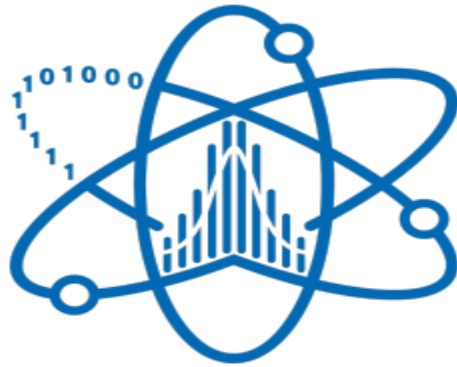








National Research
**Tomsk
State
University**



**Лаборатория
анализа данных
физики высоких энергий**

Томского
государственного
университета

Thank you for your attention!!!