

A small horizontal bar with a teal segment on the left and an orange segment on the right is positioned above the title.

CMS Machine Learning Updates

LIP Summer internship 2026

26/08/2026

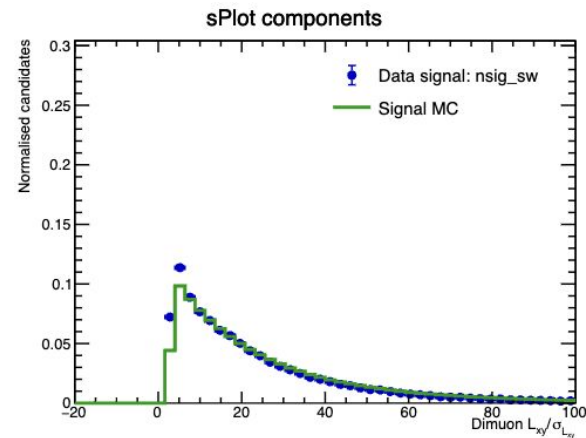
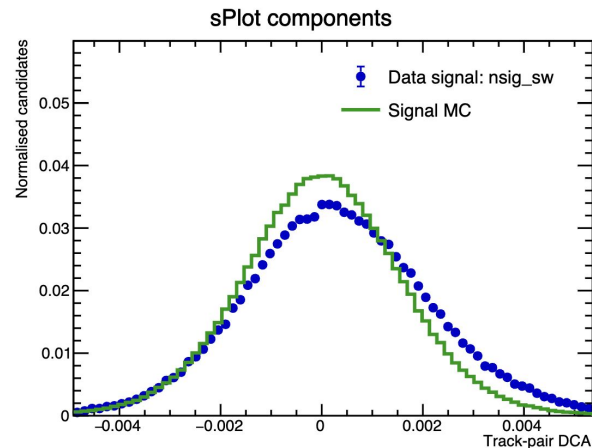
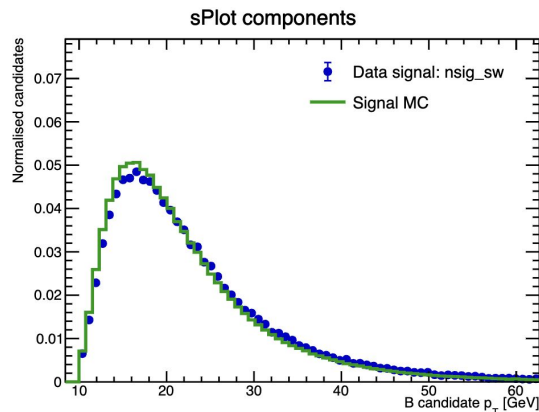
Alba Torres Soto
Bruno Ramos Veronez

Supervised by Nuno
Leonardo and Alessio Boletti

Data-MC comparison

- Clear residual shape differences visible at the one-dimensional level
- The largest discrepancy is observed in the B0 DCA-related variables, while other quantities show smaller but systematic deviations.
- These one-dimensional differences motivate a multivariate Data-vs-MC correction.

Goal: derive event-by-event weights that correct residual differences between reconstructed signal Data and signal MC.



Data-MC reweighting using XGBoost



What problems did we find?

- XGBoost rejects negative weights at input validation
- ~29% of our weights are negative

Data-MC Reweighting - Inputting Negative Weights

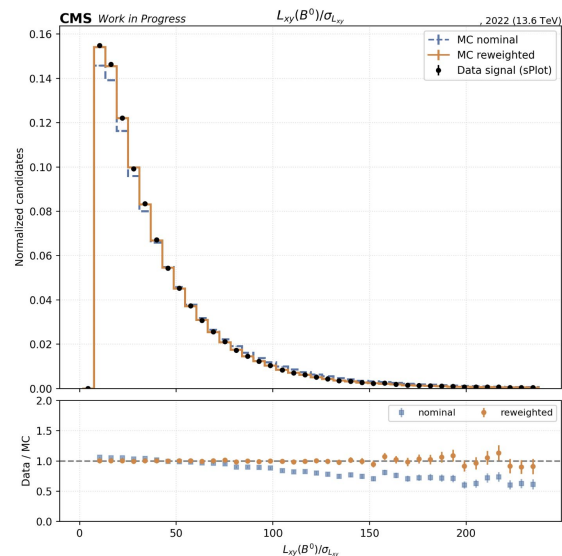
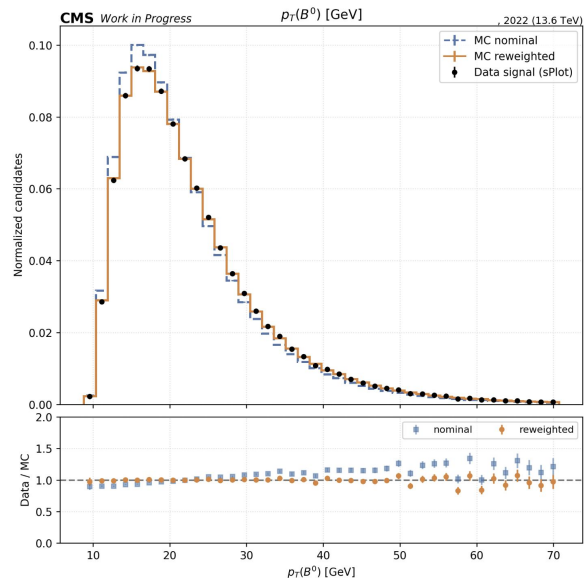
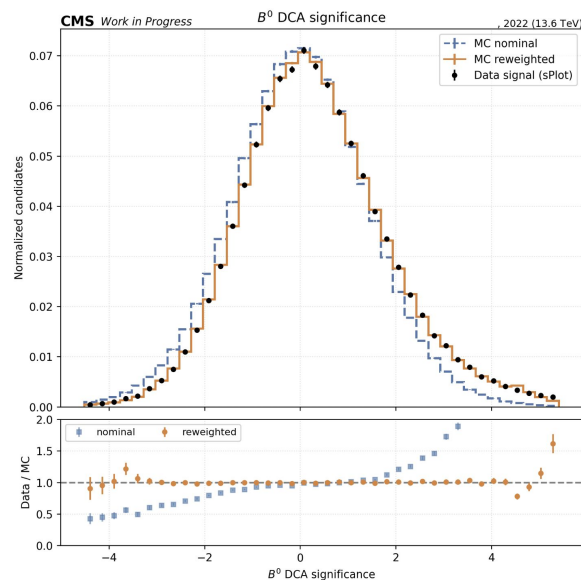


- Library swap -> **scikit-learn GradientBoostingClassifier**
- Literature ([Glazier & Tyson, arXiv:2409.08183](#)) backs it up
 - Only one class carries negative weights
 - Net weight sum positive
 - Weights enter the loss linearly
- *min_weight_fraction_leaf* = 0.01

$$\gamma_{\text{leaf}} = \frac{\sum_{i \in \text{leaf}} w_i (y_i - p_i)}{\sum_{i \in \text{leaf}} w_i p_i (1 - p_i)}$$

Data-MC Reweighting - Inputting Negative Weights

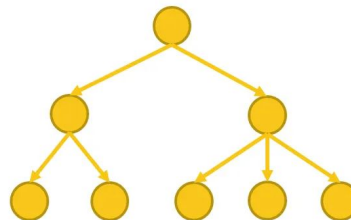
The reweighted MC is compared with the sWeighted Data signal using the original signed sWeights.



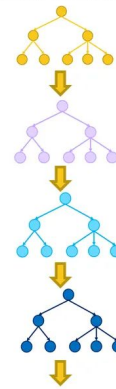
Data-MC reweighting with HEP_ML GBReweighter

- Target Data signal includes the full signed sWeights (~29% negative)
- Not a standard binary classification problem but a reweighting problem
- Using hep_ml tool: Gradient Boosted Reweighter (GBReweighter)
- **Training variables:** $L_{xy}/\sigma L_{xy}$, DCA sig., p_T , η , $\cos\alpha_{2D}$, secondary vertex probability
- Implemented with 5-fold FoldingReweighter
- Final statistics: $N_{\text{eff}}/N=90.1\%$
- Validation based on: closure plots, independent holdout tests, multi-seed robustness, weight stability

Single Decision Tree

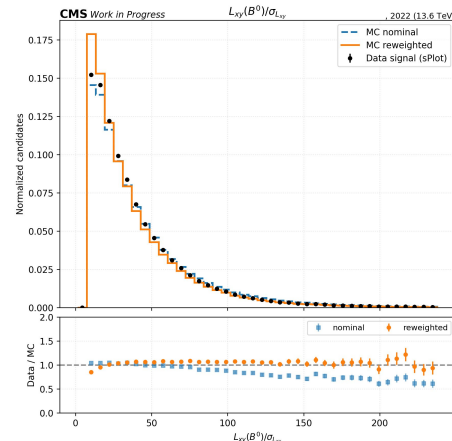
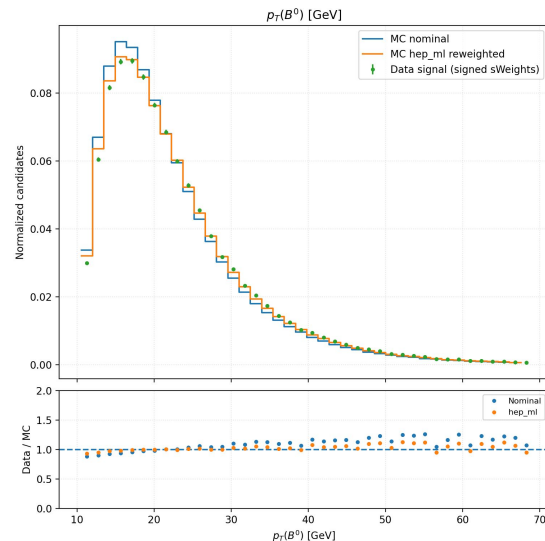
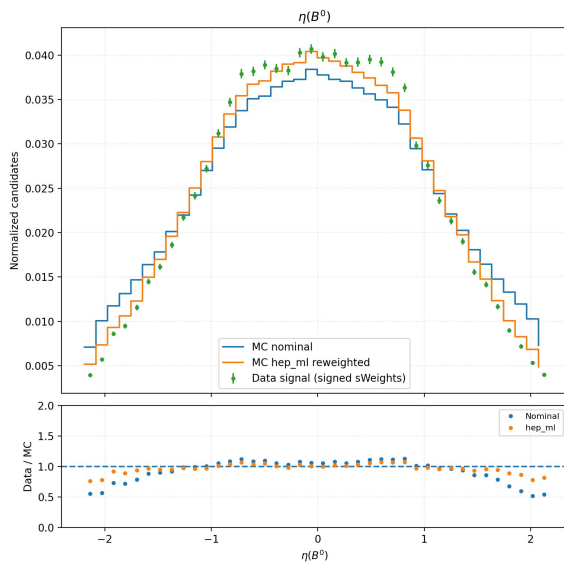
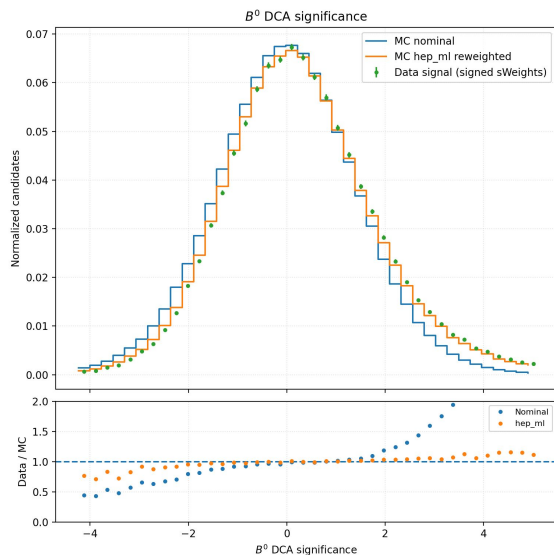


Gradient Boosted Trees



Data-MC reweighting with the

The reweighted MC is compared with the sWeighted Data signal using the original signed sWeights.



Next Steps



- Validate the reweighting on observables not used during training, e.g. individual track and muon p_T , track DCA variables
- Validate the method in the $\psi(2S)$ resonant channel
- Apply the reweighting (without retraining) to the non-resonant channel