UPDATE ON $B_s^0 \rightarrow K_S^0 K_S^0$ WITH RUN3 : BDT FEATURE OPTIMIZATION PROCEDURE

JÜRGEN F. DIVERCHY

02/09/2026

Laboratoire de Physique de Clermont
Particules pLasmas Univers applicationS
Irène Joliot-Curie Laboratoire de Physique des 2 Infinis

OBSERVABLES

- To improve signal-background separation we construct different composite variables :

$$x_{sum} = x_1 + x_2, \quad x_{max} = max(x_1, x_2), \quad x_{ratio} = \frac{x_1}{x_2}, \quad x_{asym} = \frac{x_1 - x_2}{x_1 + x_2} \text{ etc}$$

- This is motivated by the fact that axis-aligned tree splits correlations via rectangular approximations, so engineering directly the features allows single-node decisions on phase space boundaries
- We thus expanded candidate variables from the baseline (previous BDT) to 118 features

STEP 1 : SINGLE VARIABLE SEPARATION

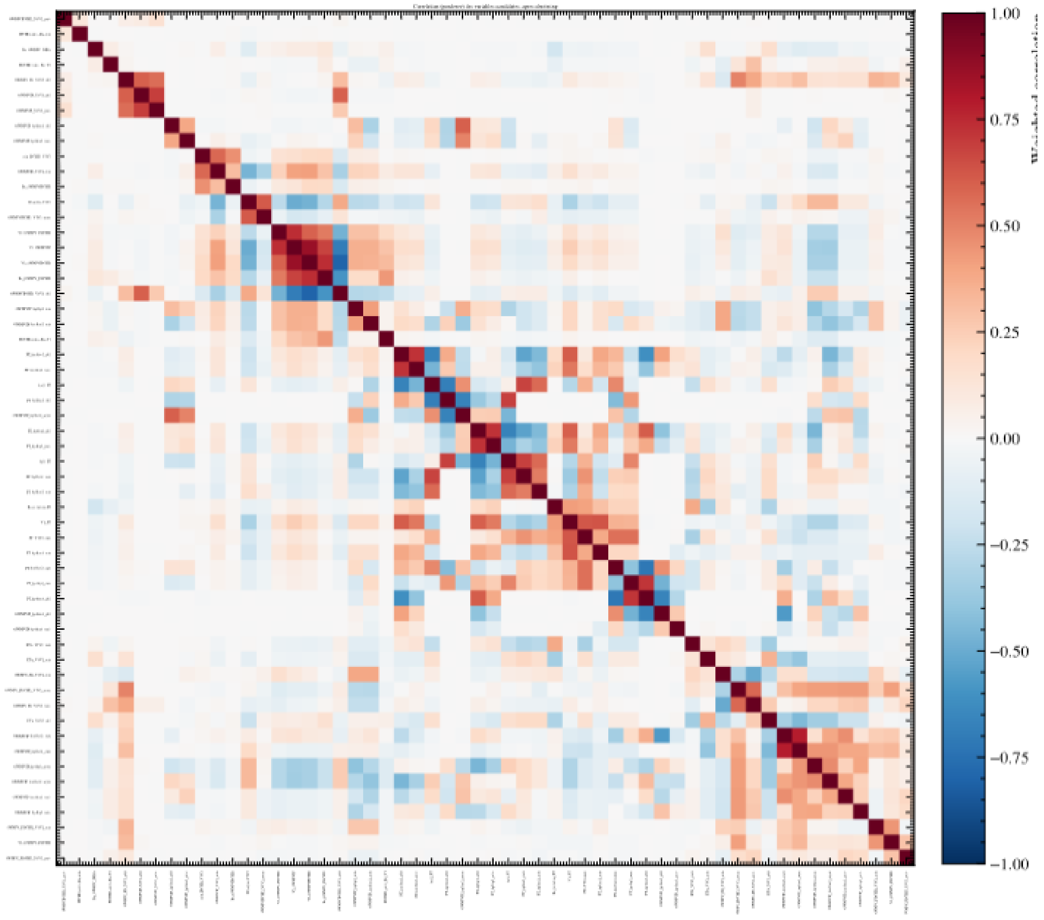
- We build the (weighted for MC) Empirical Cumulative Distribution Function (ECDF) for signal and background :

$$F_w(t) = \frac{\sum_{i=1}^{N} w_i \cdot \mathbb{I}(x_i \leq t)}{\sum_{i=1}^{N} w_i}$$

Separation capability is evaluated using the weighted two sample KS statistic :

$$D_j = \sup \left| F_{w,j}^{(S)}(t) - F_{w,j}^{(B)}(t) \right|$$

STEP 2 : REDUNDANCY REDUCTION



- To eliminate collinearity across the 118 variables, pairwise weighted Pearson correlations ρ_{jk} are calculated over the dataset to identify redundant variables. Agglomeative clustering groups variables exceeding the correlation threshold of 70% and within each cluster, only the variable with the max KS distance is kept.
- This reduced the feature space from 118 to 51 variables.

STEP 3 : RECURSIVE SHAP FEATURE ELIMINATION

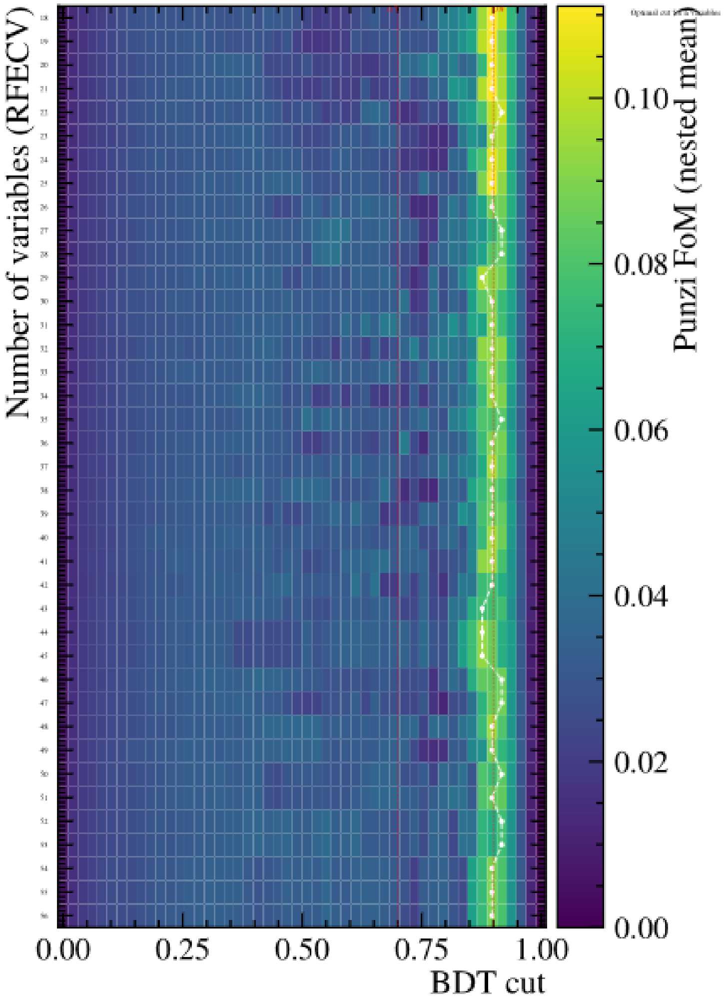
To optimize the remaining variables we perform a recursive loop of XGBoost at different depths.

At each iteration, the importance I_j of variable j is calculated across N events thanks to Shapley Additive Explanations (SHAP) :

$$I_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_{i,j}|$$

The SHAP value measures how much the j-th feature pushes event i's prediction away from the baseline value. The feature with the minimum I_j is dropped and the model is trained iteratively

STEP 3 : RECURSIVE SHAP FEATURE ELIMINATION



- For each feature count, we evaluate the optimal BDT response threshold using the Punzi FoM (with $a = 3\sigma$ significance target)

- Estimated background events in the signal mass window are extrapolated from the RHSB using a unbinned fit model

- For each feature multiplicity, the BDT threshold is scanned to find the maximum FoM value

BOOTSTRAP STABILITY ANALYSIS

To ensure robustness against statistical fluctuations in the training dataset, an independent selection is performed via bootstrap resampling (= resampling with replacement across N_{boot} iterations.

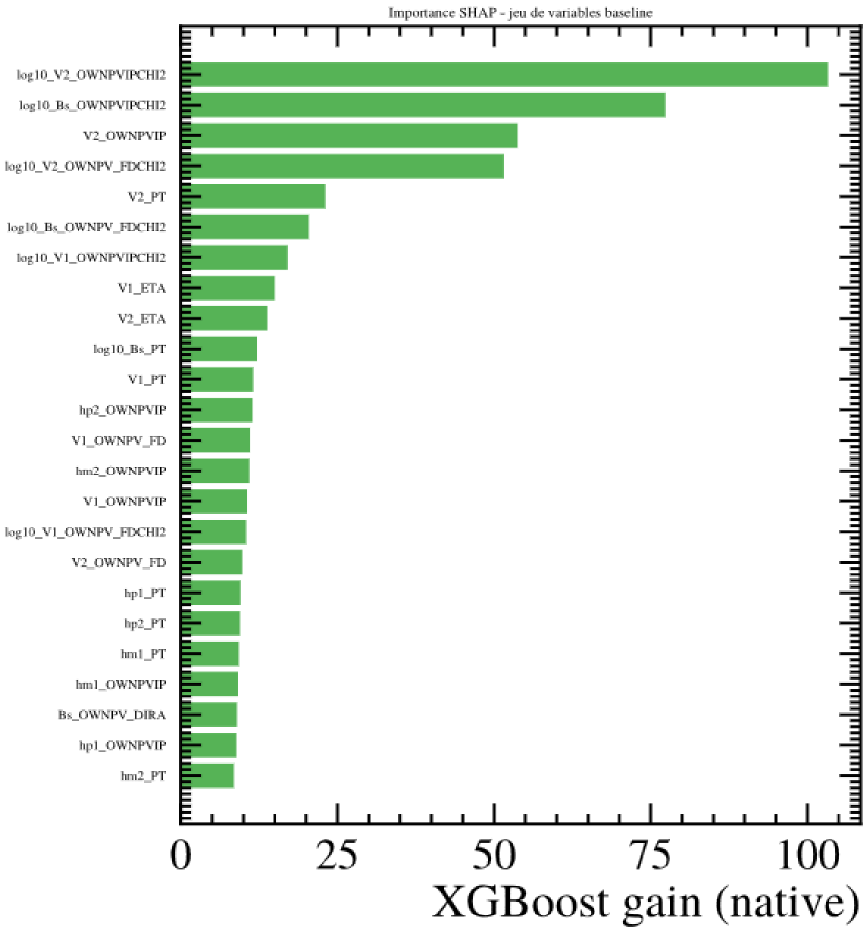
For each bootstrap replica an XGB is trained on the resampled data and features are ranked by mean absolute SHAP value. The stability score for variable j is :

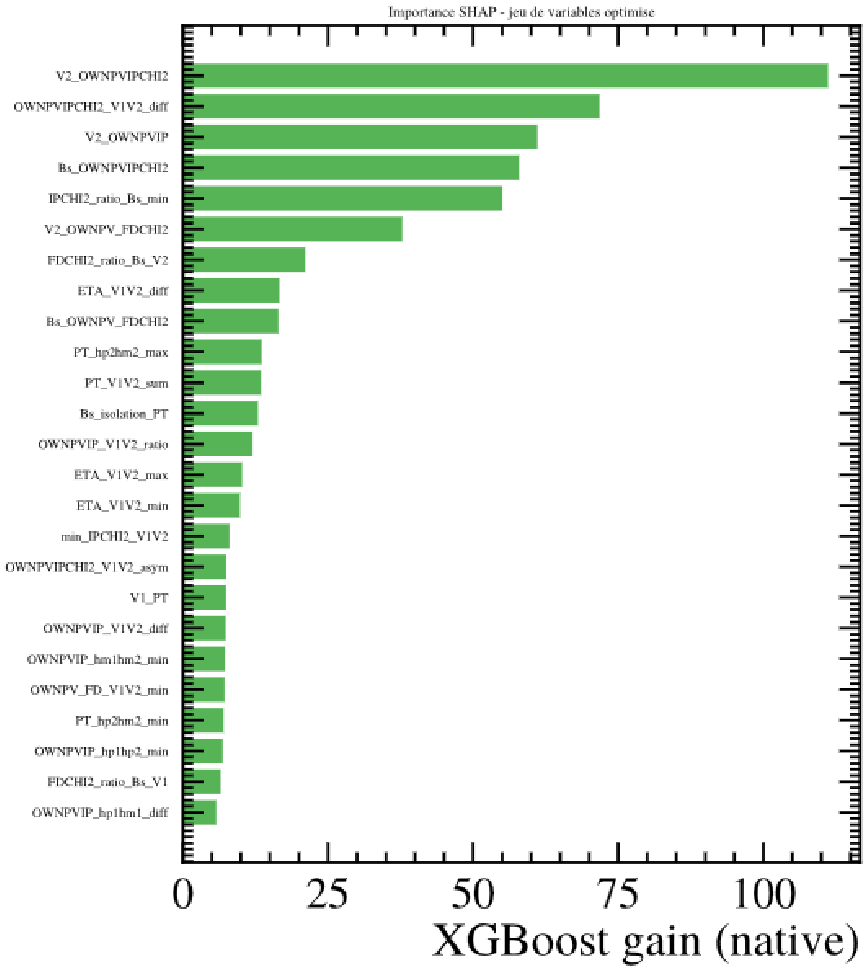
$$S_j = \frac{1}{N_{boot}} \sum_{b=1}^{N_{boot}} \mathbb{I}\left(j \in \mathcal{T}_b\right)$$

, with $\mathcal{T}_b$ the set of top-k leading variables in iteration b.

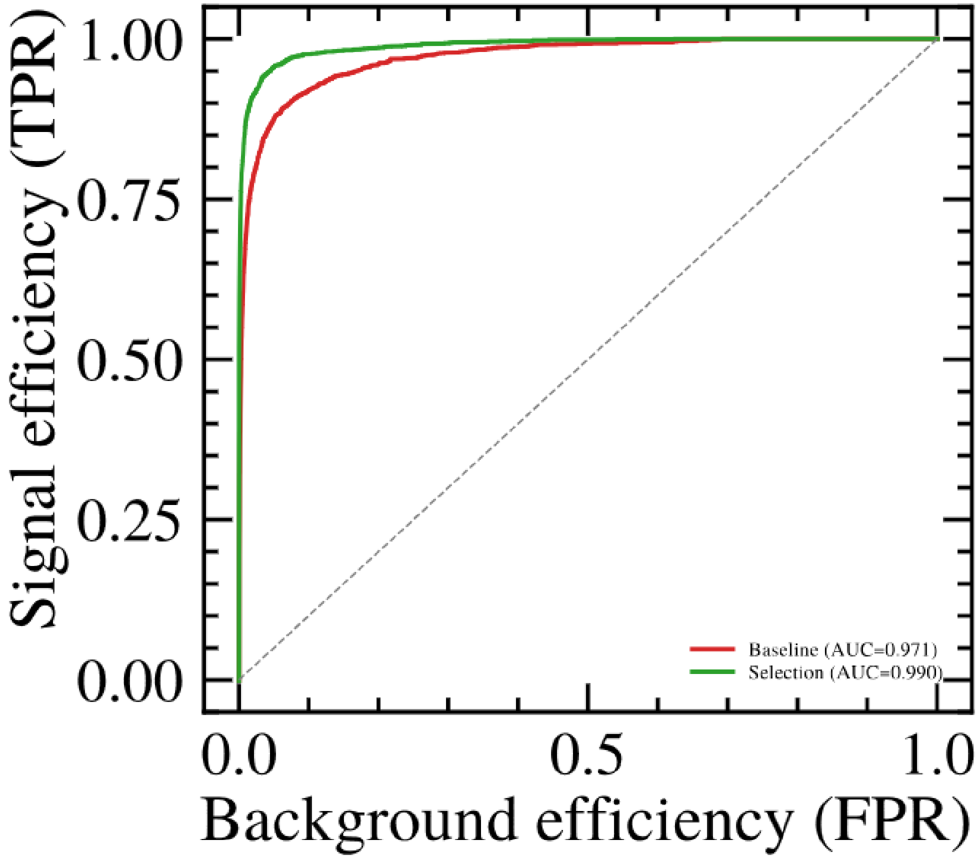
The final feature set is the intersection between the results of the two procedures and yields a compact stable set of 25 variables.

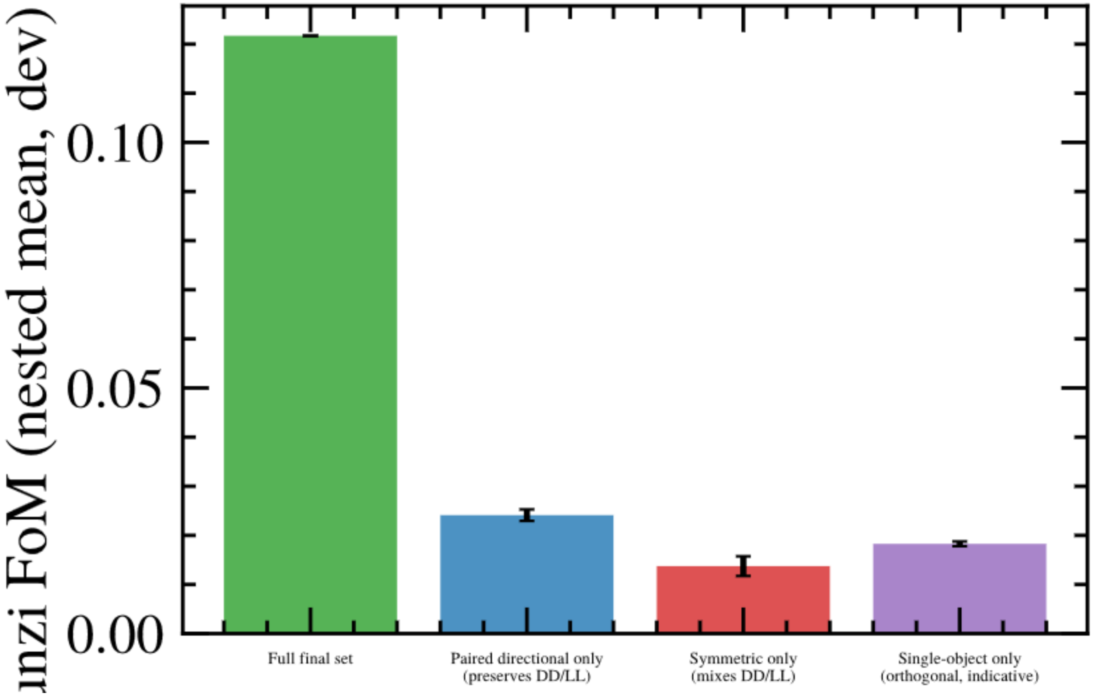
FINAL SET OF VARIABLES





BASELINE VS SELECTED COMPARISON





FUTURE PROSPECTS

- The procedure seems pretty successful
  - Also checked that the BDT is stable across depths
- Need to apply it to the LLDD sample to confirm clearly this and apply the procedure to the LLLL sample
- 2026 dataset are produced and this method could also be applied to it when the MC will be around (maybe will launch a first test with the 2025 MC)
- Still waiting for part-reco MC productions (asked liaisons for the possibility to upgrade the priority in DIRAC)
- Future refinement of the procedure may be needed (especially concerning the stability of the fits for the FoM + signal overtraining problem)