Flavour Anomalies in B-meson decays

LIP Summer internship 2026

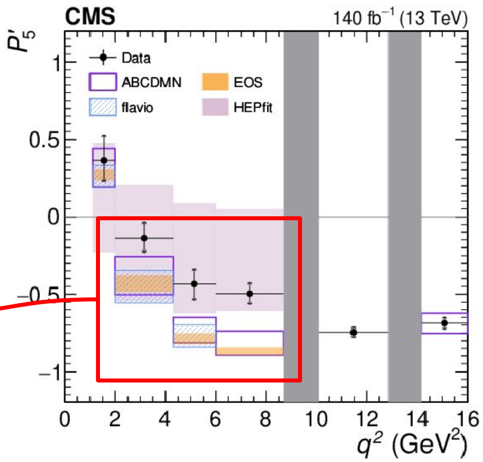
03/09/2026

Alba Torres Soto
Bruno Ramos Veronez

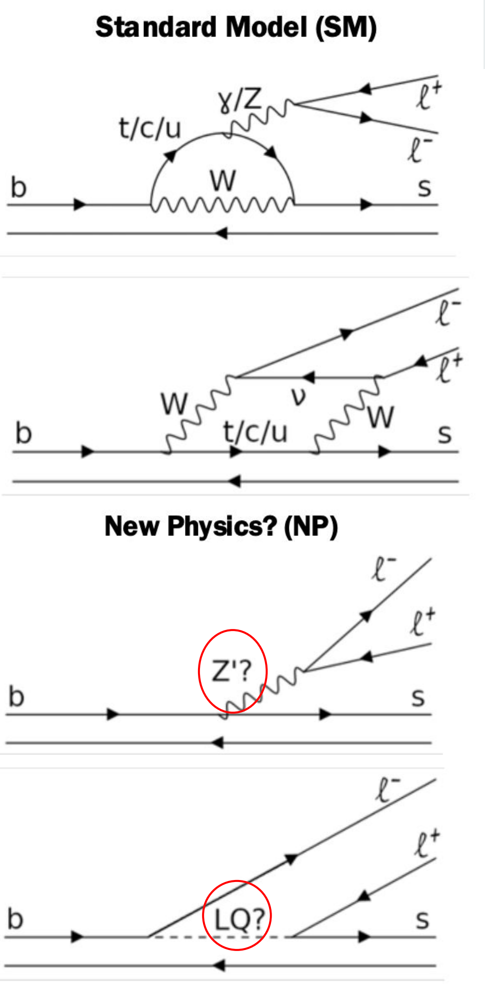
Supervised by Nuno Leonardo and Alessio Boletti

Motivation: Why study b→sμ⁺μ⁻ decays?

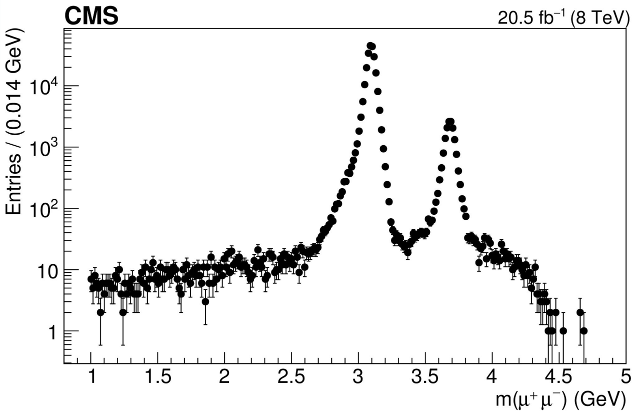
- The $B^0 \to K^{*0}\mu^+\mu^-$ decay proceeds through the rare flavour-changing transition $b \to s\mu^+\mu^-$
- This transition is strongly suppressed in the Standard Model (SM), making it sensitive to possible New Physics contributions such as a Z′ boson or a leptoquark (LQ)
- Previous CMS measurements of angular observables in this channel showed tensions with some Standard Model predictions



CMS previous analysis observed tension with some Standard Model predictions



Motivation: The decay



DECAY CHANNEL:

Non-resonant channel:

$$B^0 \to K^{*0}\mu^+\mu^- \to K^+\pi^-\mu^+\mu^-$$

Sensitive to variations of the SM

Resonant channel:

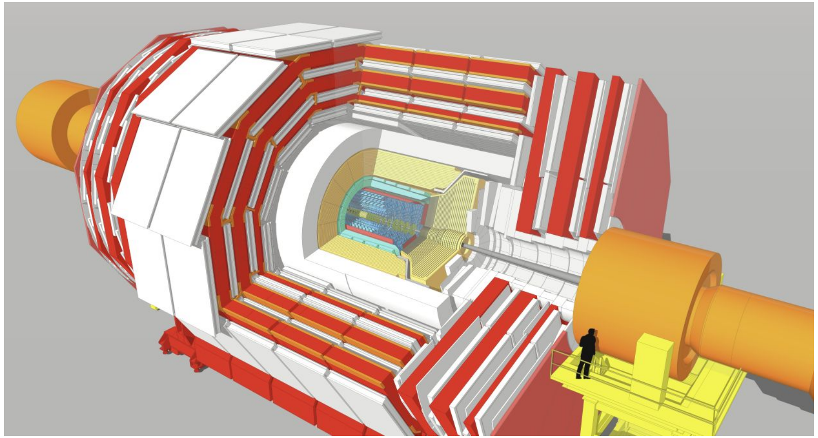
$$B^0 \to K^{*0}J/\psi \to K^+\pi^-\mu^+\mu^-$$

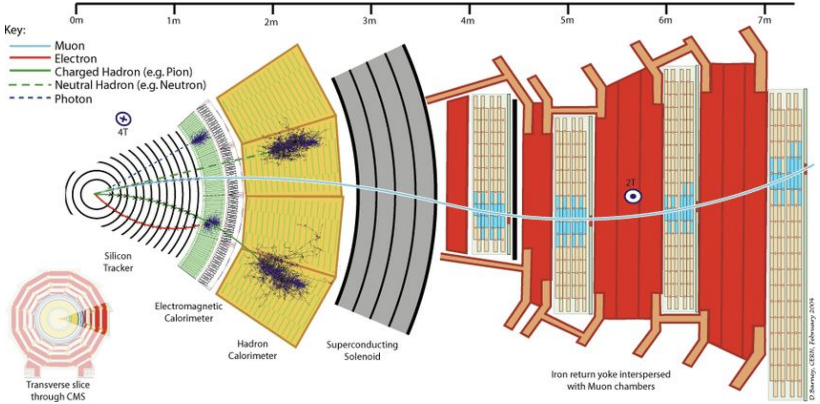
Large-statistics control region used to validate the analysis

- The resonant J/ψ channel provides a high-statistics control sample in which the analysis strategy and the Data–MC modelling can be validated.
- The validated corrections can then be transferred to the non-resonant channel, where the final physics observables are measured.

The CMS detector

- General-purpose detector with several subdetectors
- Cylindrical shape:
  - ~15m in diameter
  - ~21m in length
- **Silicon trackers** in a 3.8 T field → momentum, decay vertex
- **Muon chambers** → the $\mu^+\mu^-$ pair (triggers down to very low momentum)
- A two-level trigger selects which events to record, in real time
- Event reconstruction combines all subdetector signals into particle candidates





Analysis strategy

Data preparation
- Visualization of variables on histograms
- Build a clean J/ψ control sample
- Prepare the candidates for the mass fit

Mass fitting
- B^0 candidate mass spectrum is critical to distinguish signal from background
- Building a model to fit the mass spectrum

sPlot signal extraction
- Method to build variables' distributions of signal in data
- Compare sWeighted signal Data with signal MC

Data-MC reweighting
- Derive event-by-event MC correction weights
- Two different Machine Learning approaches

Final result
- Apply corrections to all MC samples

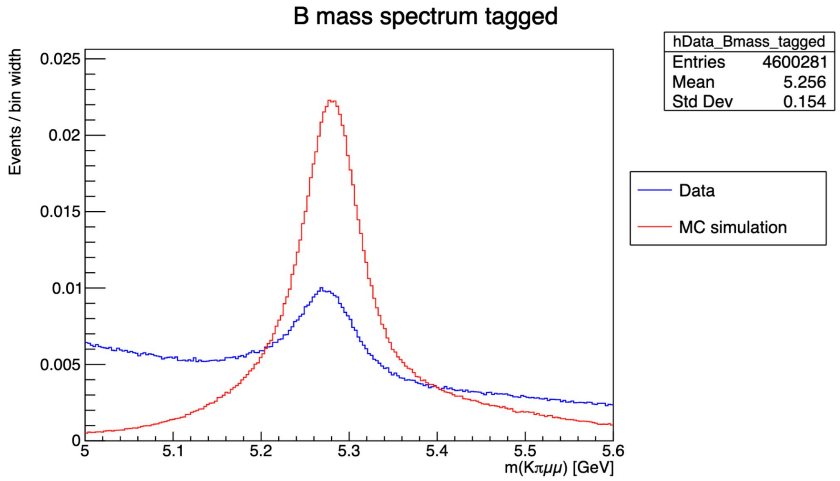
Long-term CMS analysis goal: measure the physics observables in the non-resonant $B0 \rightarrow K*0 \mu+ \mu-$ channel. The corrections developed in this project will be integrated into the full analysis workflow.

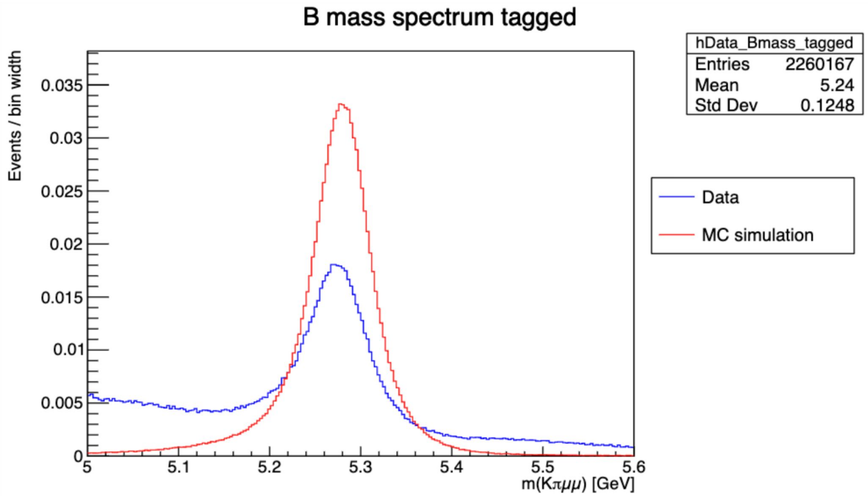
Data preparation

<u>Data set:</u> **Run 3 of the CMS experiment (2022) - pp collisions (13.6 TeV)**

- **Preselection cuts** on topological and kinematic variables
- **Single candidate selection:** multiple candidates per event initially $\rightarrow$ selection based on best vertex probability
- **Veto applied** to reject $B_s \rightarrow \phi\mu^+\mu^- \rightarrow K^+K^-\mu^+\mu^-$ peaking background
- **Flavour tagging** applied since CMS detector cannot distinguish between kaons and pions

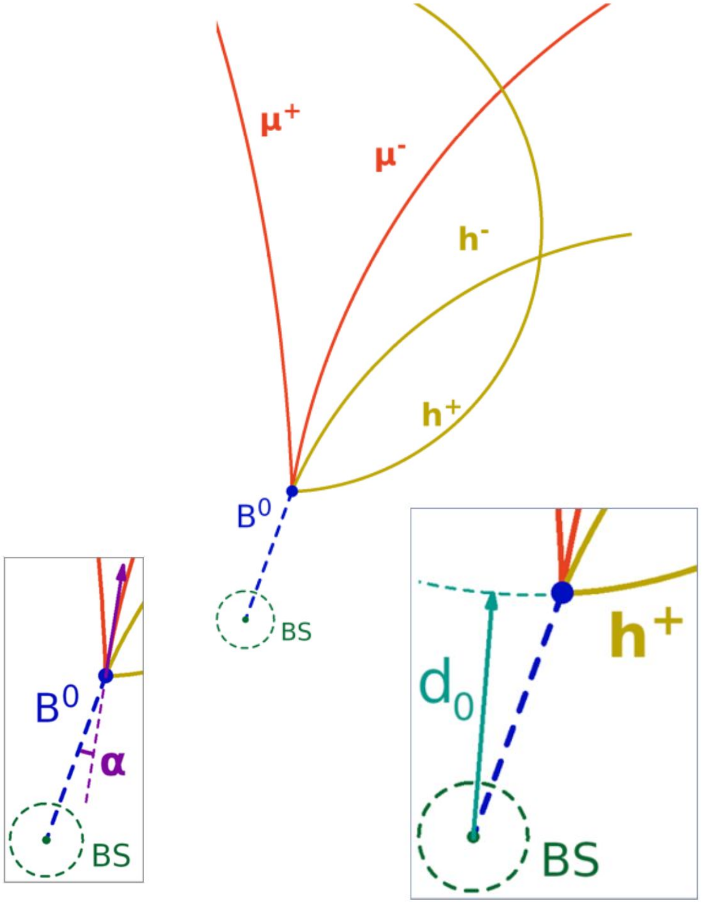
$$B^0 \rightarrow K^{*0}\mu^+\mu^- \rightarrow K^+\pi^-\mu^+\mu^-$$
$$\bar{B}^0 \rightarrow \bar{K}^{*0}\mu^+\mu^- \rightarrow K^-\pi^+\mu^+\mu^-$$





Key variables

- **Transverse Momentum (pT):** momentum component perpendicular to the LHC beams.
- **Pseudorapidity (η):** describes the polar angle of a particle relative to the beam line
- **Distance of Closest Approach (DCA):** minimum distance between a particle's trajectory and the collision region on the transverse plane
- **Secondary Vertex Probability (svprob):** fitted probability that candidate daughter tracks (π,K,μ+,μ−) originate from a single, well-defined decay point.
- **cos α:** angle between the reconstructed B0 momentum and the line connecting the Primary Vertex (PV) and Secondary Vertex (SV)
- **Flight Distance Significance (Lxy/σLxy):** transverse distance between the Primary Vertex and Secondary Vertex, normalized by its uncertainty.



Why We Fit The Mass?

- Mass is a critical variable to separate signal and combinatorial background:
  - Signal peaks
  - Background is smooth across the window
- For a fraction of candidates (~13.5%), the flavour tagging assigns the K and π masses to the wrong tracks
- **Simulation first:** RT and WT are isolated in MC and each shape is modelled individually
- **Data second:** The shapes are fixed from the MC fits. Only a small set of parameters is left free in the data fit

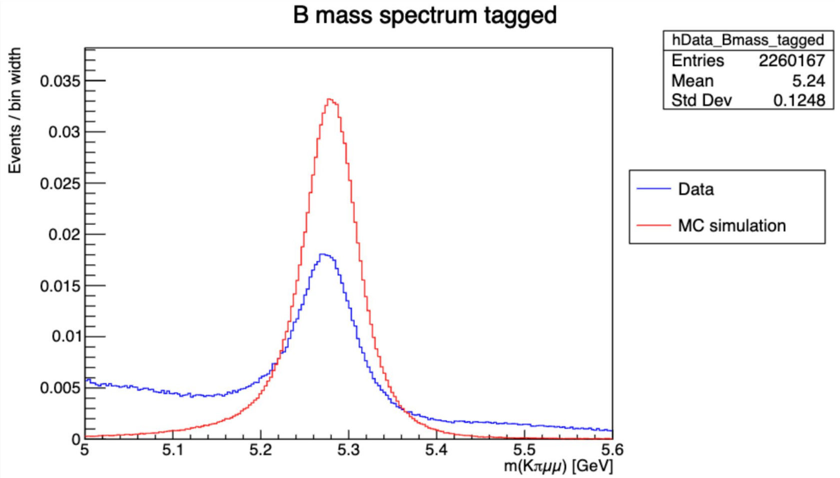
$$\mathcal{P}_{\mathrm{sig}}(m) = f_{\mathrm{RT}}\,\mathcal{P}_{\mathrm{RT}}(m) + (1 - f_{\mathrm{RT}})\,\mathcal{P}_{\mathrm{WT}}(m)$$

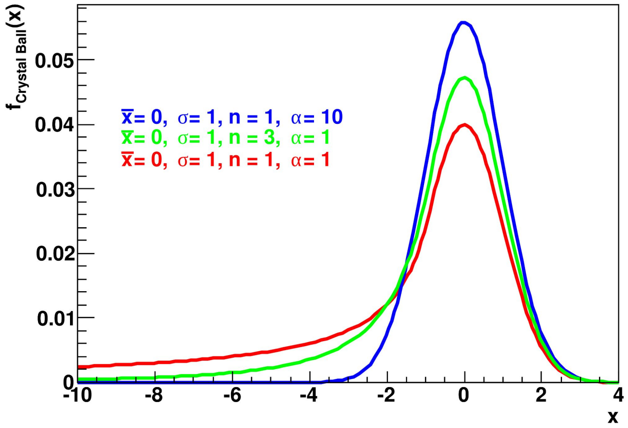
What The Signal Looks Like

- Gaussian Core
  - All components share <u>one mean</u>
- Asymmetric Tails

Modelling

- **Main function**: Double-Sided Crystal Ball
  - Gaussian core spliced to a power-law tail
- **Auxiliary functions**: additional Gaussians





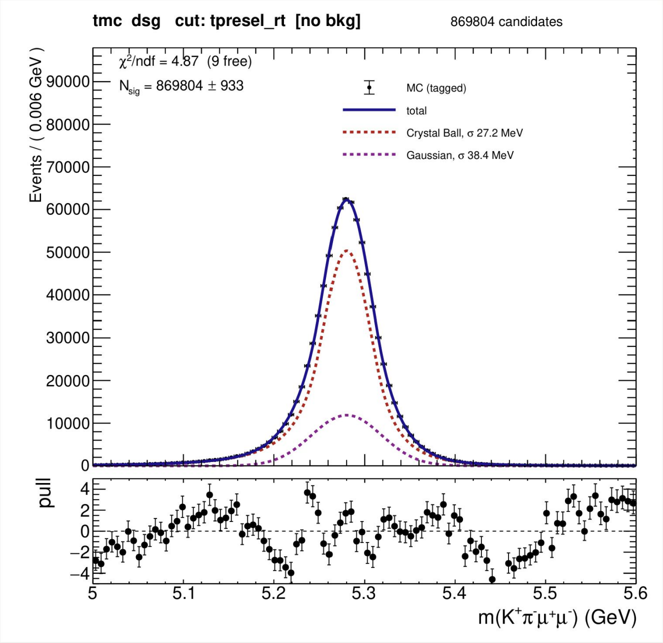
Criteria

Multiple options were tested to build the RT and WT models

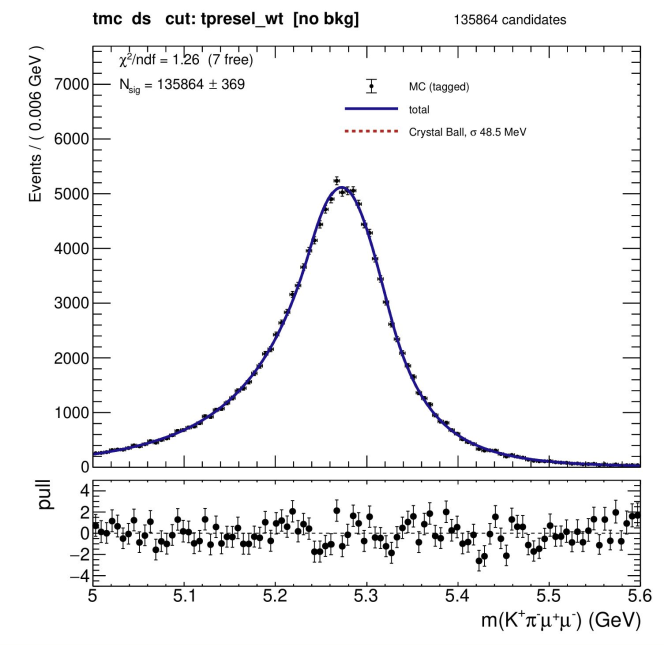
The best were chosen based on:

- **Physical validity**
- **Convergence and audit**: MIGRAD status, covariance quality, no floating parameter at a limit or unconstrained.
- **χ^2/ndf and pulls**: the final judge

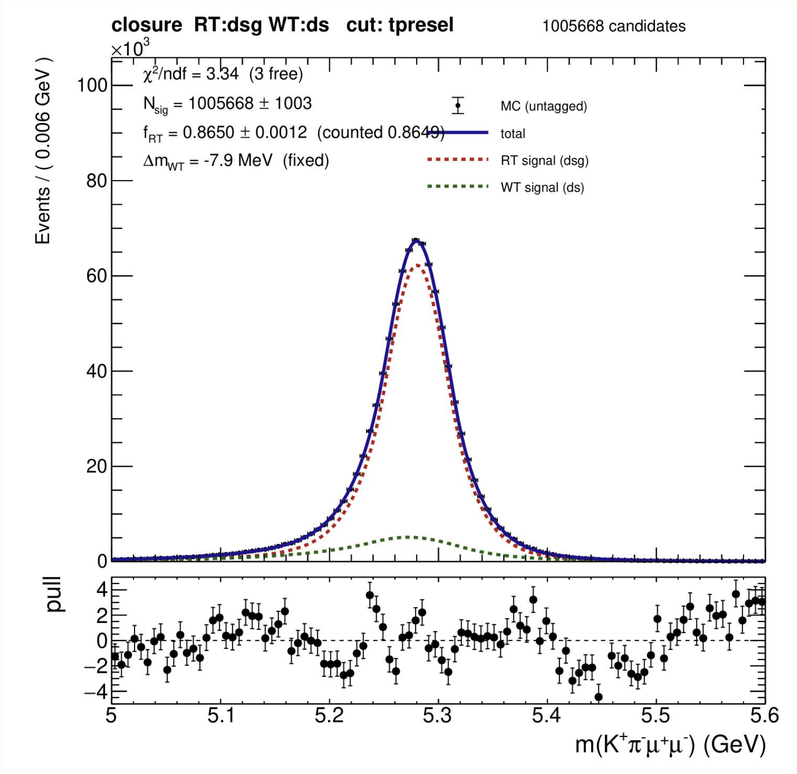
RT = Crystal Ball + Gaussian



WT = Crystal Ball



Combined Fit On MC

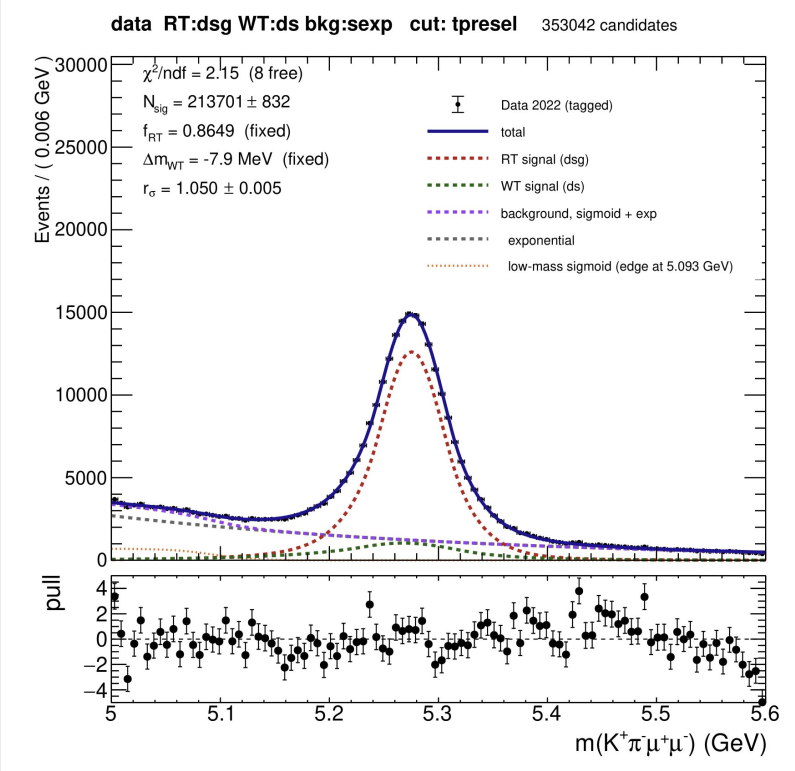


- **All shape parameters are frozen**
- WT peak is tied to RT peak by Δm_{WT}

Free Parameters:

- μ (peak position)
- f_{RT} (right-tag fraction)
- N_{sig} (signal yield)

Combined Fit On Data



$$\mathcal{P}(m) = N_{\text{sig}}\left[f_{\text{RT}}\,\mathcal{P}_{\text{RT}}(m) + (1-f_{\text{RT}})\,\mathcal{P}_{\text{WT}}(m)\right] + N_{\text{bkg}}\,\mathcal{P}_{\text{bkg}}(m)$$

Background: exponential and a sigmoid function, to follow the rise in the left sideband!

Free Parameters:

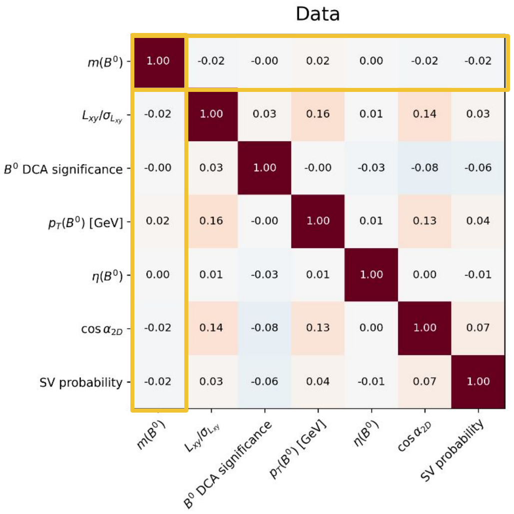
- f_{prc}
- m_0
- k
- λ

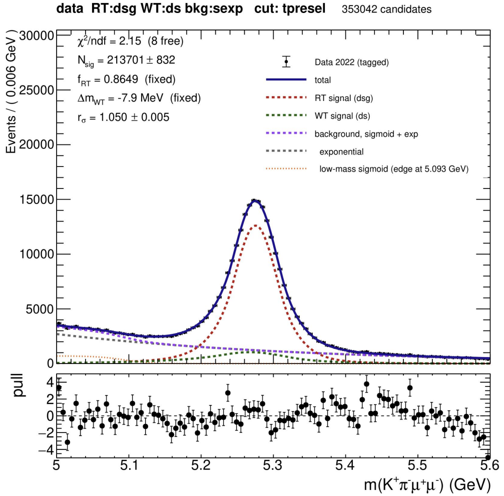
$\underbrace{\hphantom{f_{\text{prc}}\,m_0\,k\,\lambda}}_{\text{Background Shape Related}}$

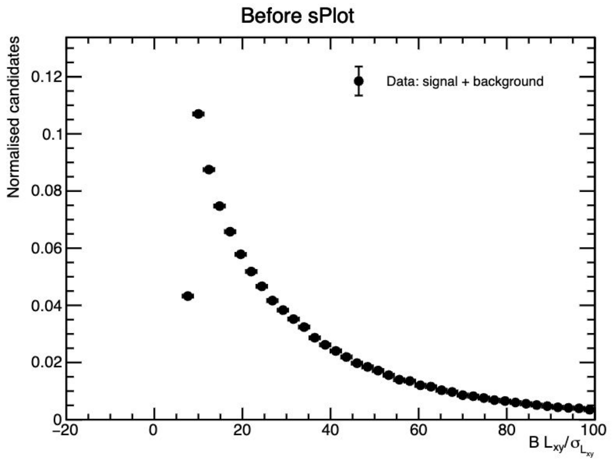
- μ (peak shift $\rightarrow$ momentum scale)
- r_σ (width multiplier $\rightarrow$ resolution)
- N_{sig}
- N_{bkg}

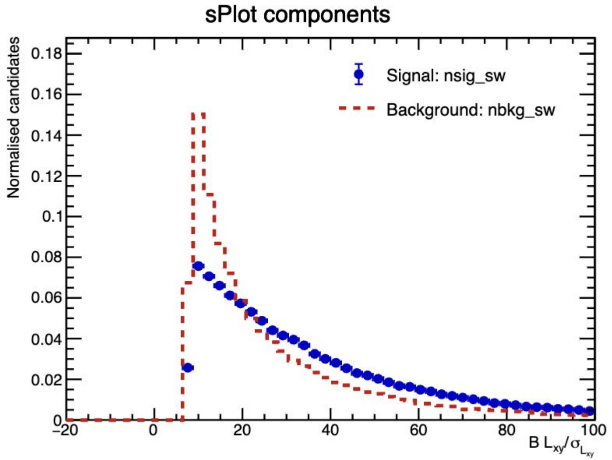
SPlot

Tool that assigns per-event **statistical weights to separate the signal and background** contributions in Data using the fitted discriminating variable (B-candidate mass in our case)



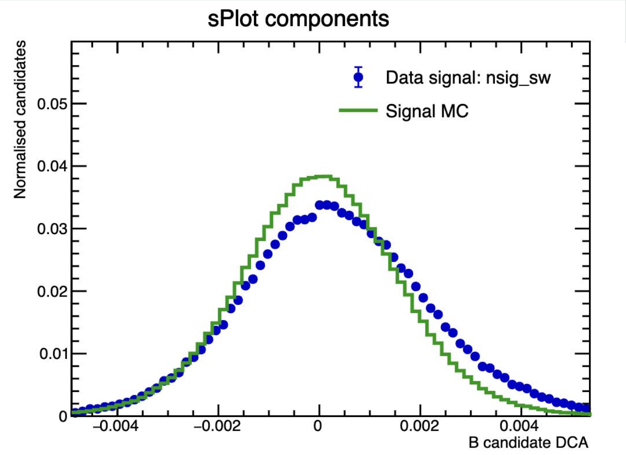




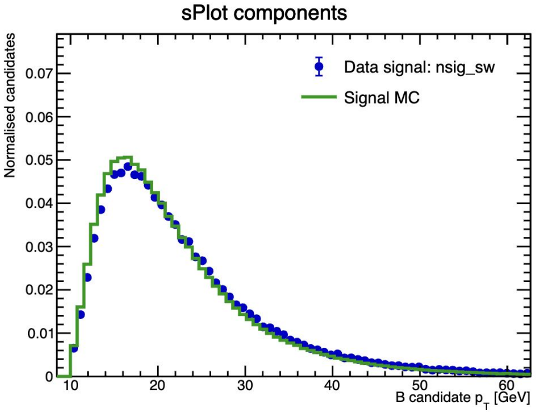


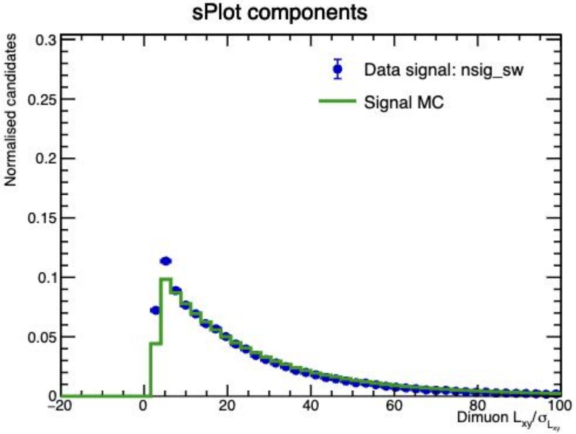
Data-MC comparison

- Comparison between signal in Data (SPlot weights) vs signal in MC
- Clear residual shape differences visible at the one-dimensional level
- 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: Machine Learning approaches

- **Goal**: derive event-by-event weights that correct residual differences between reconstructed signal Data and signal MC.
- training performed with: data with sWeights (to mimic the signal-only population) vs MC
- Target Data signal includes ~29% negative weights which is a huge drawback for many ML tools

⇩

- Not a standard binary classification problem but a reweighting problem
- **Training variables:** $L_{xy}/\sigma L_{xy}$, DCA sig., p_T, η, $\cos\alpha 2D$, secondary vertex probability
- Once the training is completed we can apply the new weights to any variable without needing to retrain

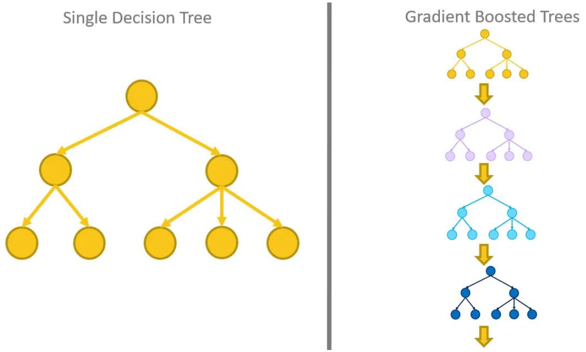
Data-MC Reweighting

Scikit library

- Using **scikit-learn GradientBoostingClassifier**
- Literature ([Glazier & Tyson, arXiv:2409.08183](https://arxiv.org/abs/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*

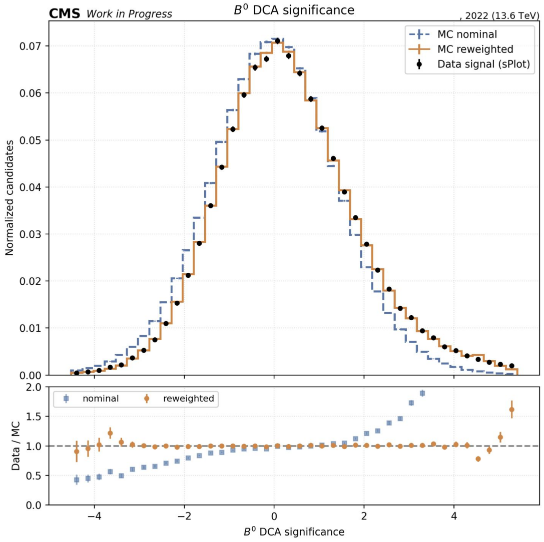
HEP_ML library

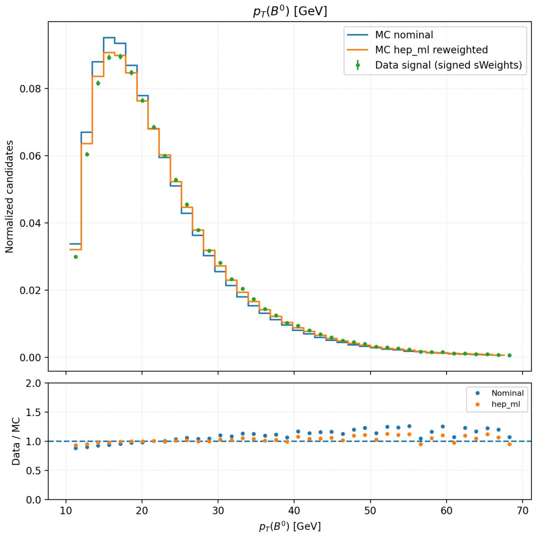
- Using hep_ml tool: **Gradient Boosted Reweighter (GBReweighter)**
- Target Data signal includes the full signed sWeights
- Final statistics: Neff/N=90.1%
- Validation based on: closure plots, independent holdout tests, multi-seed robustness, weight stability

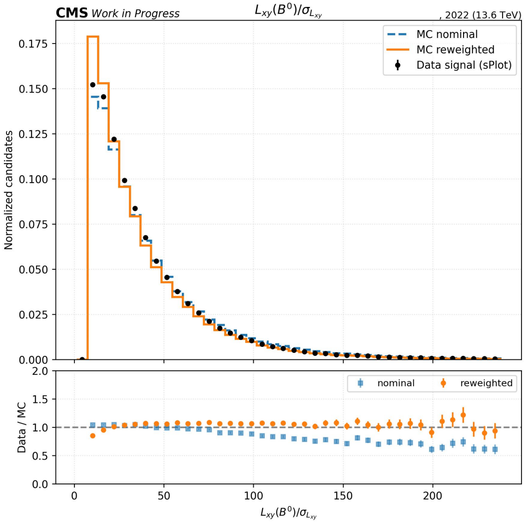


Data-MC Reweighting

After applying the reweighting, the MC distributions for all variables largely improve the compatibility with the data.







Summary

- Studied B0 meson signals in LHC Run3 data collected by CMS in pp collisions (13.6 TeV)
- Performed RT/WT mass fitting, SPlot signal extraction, and final MC reweighting with two machine learning approaches
- Data–MC correction strategy developed and validated in the J/ψ resonant control region

Next Steps

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