Centro de Astropartículas y Física de Altas Energías
Universidad Zaragoza

Departamento de Física Teórica
Universidad Zaragoza

Using Machine Learning techniques in phenomenological studies in flavour physics

Jorge Alda,
Universidad de Zaragoza/CAPA jalda@unizar.es

Based on **JA**, J. Guasch, S. Peñaranda
arXiv:2109.07405 [hep-ph]

5th Inter-experiment Machine Learning Workshop

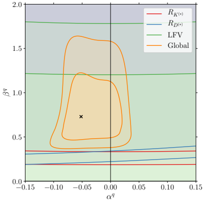
12th May 2022

Problem

- We study $R_{K^{(*)}}$ and $R_{D^{(*)}}$ anomalies affecting B meson decays
- using Effective Field Theory at $\Lambda = 1\,\mathrm{TeV}$.

$$\mathcal{L}_{\mathrm{EFT}} = \mathcal{L}_{\mathrm{SM}} + \frac{1}{\Lambda^2} C \lambda^{\ell}_{ij} \lambda^{q}_{kl} \left[(\bar{\ell}_i \gamma_\mu \ell_j)(\bar{q}_k \gamma^\mu q_l) + (\bar{\ell}_i \gamma_\mu \tau^I \ell_j)(\bar{q}_k \gamma^\mu \tau^I q_l) \right].$$

- Global fits with 5 parameters (C, α^ℓ, β^ℓ, α^q, β^q), log-likelihood function contains 471 physical observables $\Longrightarrow$ **High computation time**.
- Non-linear relations $\Longrightarrow$ equi-probability regions are not elliptical $\Longrightarrow$ **we can not use Hessian approximation for the log-likelihood**.

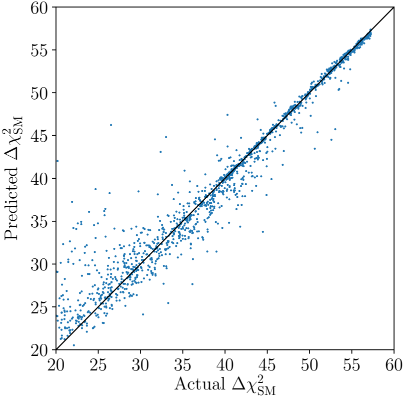


Machine Learning training

Solution: We create an approximation of the log-likelihood function using the `xgboost` model (regression tree).

Data sample consisting of

- 5000 points re-used from likelihood plots.
- 5000 random points.
- Split in 75% training set, 25% validation set.
- Learning rate 0.05, 1000 estimators, early stopping at 5 rounds.



Results of the training:

- Pearson regression coefficient $r = 0.971$.
- Mean Absolute Error 0.655.

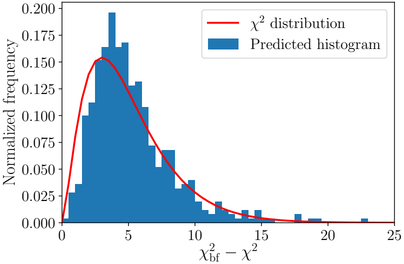
Machine Learning Montecarlo

We want to generate a sample of points in parameter space distributed according to the χ^2 of the fit.
We generate random points, that are accepted if

$$\log \tilde{L}(\vec{C}) = \log L_{\rm bf} + \log u\,,$$

with u a random number from the uniform distribution in $[0,1)$.

Montecarlo points generated using the Machine Learning algorithm:

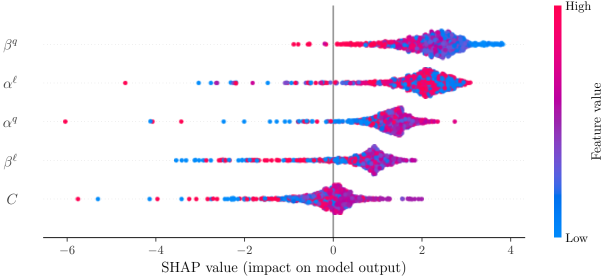


We use the trained model $\log \tilde{L}(\vec{C})$ to compute an approximation of the likelihood function.

SHAP values

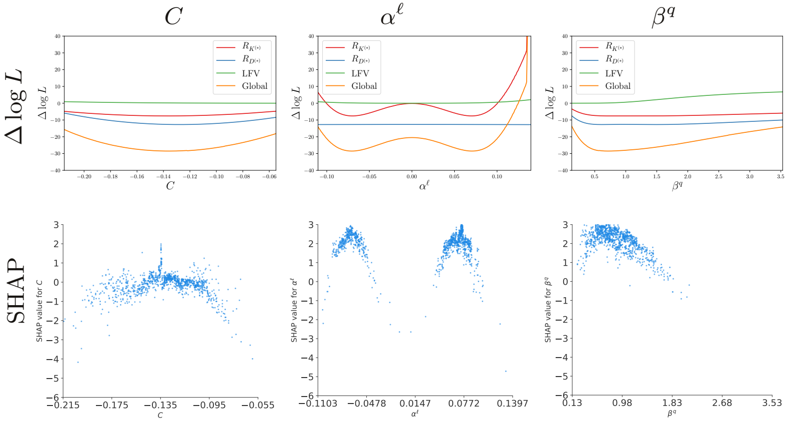
We calculate SHAP values in the Montecarlo sample to examine the importance of each parameter in the xgboost predictions.

The mixing β^q to the second quark generation ($b \to s$ and $b \to c$) and α^ℓ to the first lepton generation (explains $R_{K^{(*)}}$) are in general the most important features.



SHAP values and likelihood

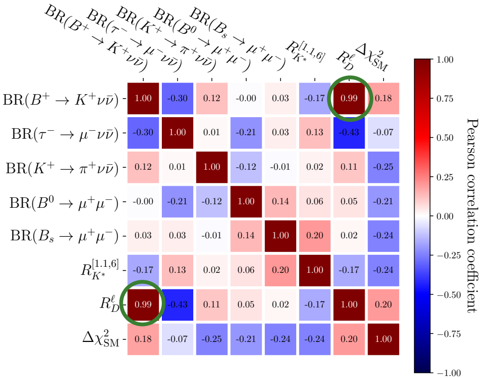
SHAP importances reproduce the dependence of the $-\log L$:



- The SHAP values reproduce correctly the general features of the fit.

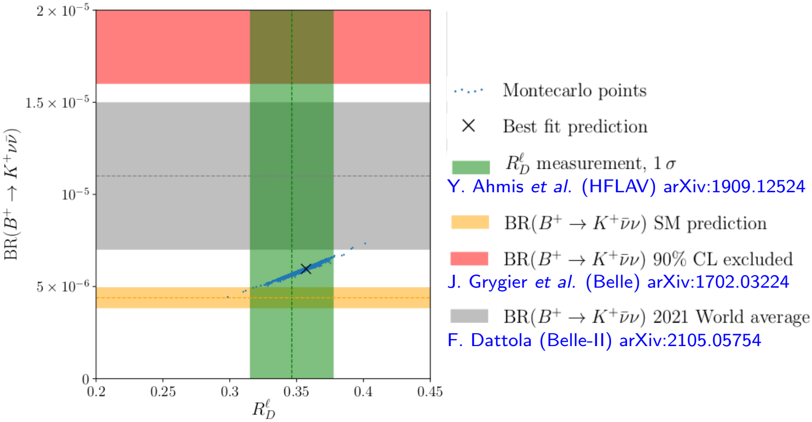
Correlations between observables

Matrix of Pearson correlation coefficients for selected observables in the Montecarlo sample:



- Moderate correlation between R_K and $\mathrm{BR}(B_s \to \mu^+\mu^-)$ because $C_9^\mu \neq C_{10}^\mu$.
- Also moderate correlation between R_K and R_D.
- Perfect correlation between R_D $\mathrm{BR}(B \to K^{(*)}\nu\bar{\nu})$.
- No observable displays large correlations to the global likelihood: global fits are needed.

R_D and $\mathrm{BR}(B \to K^{(*)}\nu\bar{\nu})$



An excess in R_D implies an excess in $\mathrm{BR}(B \to K^{(*)}\nu\bar{\nu})$.
(Note that the 2021 World Average is not included in our fit).

Conclusions

We have applied Machine Learning techniques to the flavour phenomenology of B anomalies.

- We have trained an approximation of the global log-likelihood function using `xgboost`.
- We can generate new samples using a Montecarlo based on the ML approximation.
- We have identified the most important parameters using `SHAP` values.
- We have studied the correlations between physical observables and compared them to experimental results.

More info at arXiv:2109.07405 [hep-ph]
Code at `https://github.com/Jorge-Alda/SMEFT19`