B-meson angular observable fitting using Google Tensorflow

L. Shepherd

A thesis submitted in fulfilment of the requirements for the degree of Master of Physics and the Diploma of Imperial College London

September 23, 2019

Abstract

Angular observables of $B^0 \to K^{*0}\mu^+\mu^-$ decays have been found to be in tension with the standard model. This study builds on a previously proposed method for performing fits to $B^0 \to K^{*0}\mu^+\mu^-$ angular observables. In this work, Google Tensorflow is used against a generated toy signal. It was found that Tensorflow could perform the fits, but the results are dependant on machine learning hyperparameters and the coefficient initialisation algorithm. A 4.9 σ confidence level is found for differentiating the standard model from a new physics model using 600 signal events.

Contents

1 Introduction

Various measurements for B decays have been found to be in tension with the standard model (SM). The SM predicts lepton-universality where $B \to K\ell\ell$ decays should proceed equally to $B \to Kee$ and $B \to K\mu\mu$, so the branching ratio would be expected to be unity. However what has been found is [1]

$$R_K = \frac{\mathrm{BR}(B \to K\mu\mu)}{\mathrm{BR}(B \to Kee)} = 0.846\ ^{+0.060}_{-0.054}\ ^{+0.016}_{-0.014}, \qquad \text{for } 1.1\,\mathrm{GeV}^2 < q^2 < 6\,\mathrm{GeV}^2, \quad (1)$$

where q^2 is the invariant lepton mass squared. The uncertainties correspond to statistical and systematic respectively. This result represents a 2.5 σ tension. The sensitivity of this result is only currently restricted by statistics with theoretical uncertainties negligible.

The SM only allows flavour-changing neutral current (FCNC) interactions such as $B^0 \to K^{*0}\mu^+\mu^-$ via loop decays (see Fig. 1), disallowing tree-level decays. The loop structure of these decays could have contributions from new heavy particles which would alter the branching ratios and angular observables. These decays are therefore of interest in to the search for new physics (NP) [2].

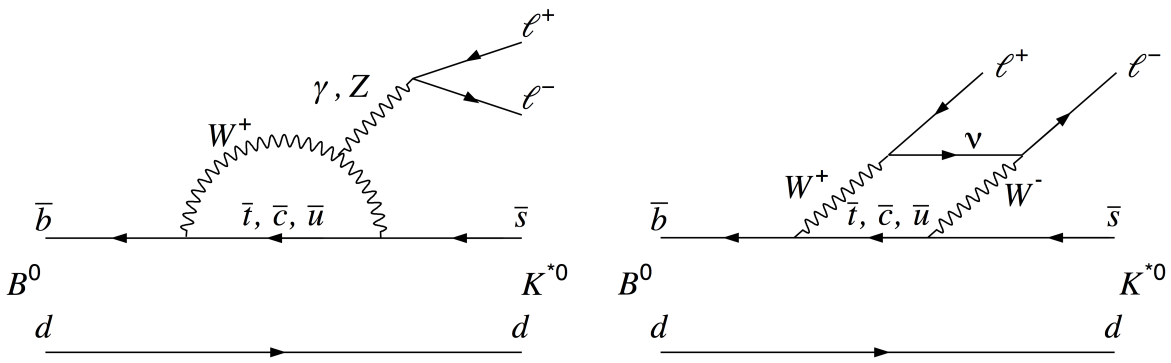

Figure 1: SM Feynman loop diagrams for the $B^0 \to K^{*0}\mu^+\mu^-$ decay via $b \to s$ electroweak penguin decays (left) and box decays (right). [2]

An analysis of the angular observables of the $B^0 \to K^{*0}\mu^+\mu^-$ decay was performed at the LHCb experiment and found to be in 3.4 σ tension with the SM [3]. This decay is attractive to study as the energy of the kaon makes it self-tagging. Unlike in the branching ratio case, sensitivity to the angular observables is restricted by theoretical uncertainties associated with hadronic form factors. Short distance factors can be computed for $q^2 < 10\,\mathrm{GeV}^2/c^4$ using Light-Cone Sum Rules (LCSRs) via a QCD operator product expansion [4]. However lattice QCD must be used for higher energies [5]. There are also long distance factors that can be computed using perturbation theory at lower energies where $q^2 \ll 4m_c^2$ ($\sim 15\,\mathrm{GeV}^2/c^4$) with m_c the mass of the charm quark. At higher energies they are subject to a "charm-loop effect" where more complicated treatments are required [5]. Working in the low energy regime therefore allows the uncertainties

to be reduced. Furthermore, by choosing to work with suitable observables that reduce the reliance on the form factors uncertainties can be reduced further. In particular the so-called $P_i^{(\prime)}$ basis does this by maximising cancellations of the form factors [6, 7].

Modified Wilson coefficients (WCs) have been proposed that could explain these tensions. Shifting the single combination $\mathcal{C}_9 = -\mathcal{C}_{10}$ by -15% of the SM value provides a 6 σ pull to the observed data. Making the dual modification $\mathcal{C}_9 \approx -0.73$ and $\mathcal{C}_{10} \approx 0.40$ shows a pull of 6.3 σ, but slight tension still remains with the branching ratio measurement. Furthermore by splitting the WCs into a lepton flavour universality component and a purely muonic component, a 6.5 σ pull is observed. It is possible that a U_1 leptoquark or two new scalar bosons could explain these results [1]. Placing further constraints on the WCs is therefore important to motivating searches for NP.

Previous fits to the observables for $B^0 \to K^{*0}\mu^+\mu^-$ have relied on a q^2 binned approach, which introduces additional uncertainties. A novel method was proposed by Egede, Patel and Petridis in 2015 that removed this binning by approximating the spin amplitudes as a three parameter anzatz that was dependent on a continuous q^2 parameter [8]. The range $1\,\mathrm{GeV}^2/c^4 < q^2 < 6\,\mathrm{GeV}^2/c^4$ was used which reduced uncertainties from the form factors as previously discussed. The method also explicitly included the S-wave components whereas previously they had been factored in by a systematic uncertainty. Based on a toy signal with parameters found through the EOS program and a 600 signal count based on Run-I results, a 6.5 σ ability to differentiate the SM from a NP model with modified WCs and CP-averaged observables was found. A three binned comparison was only able to achieve 5.0 σ. The new method represented a 30% increase in sensitivity. The model was unable to converge to a fit on more noisy real world data however.

This study implements the method set out in the previous study, but using Google's Tensorflow software. Tensorflow is a general purpose computational framework whose special focus on machine learning will be leveraged in this study. Machine learning has been previously found to be useful for B-meson flavour detection [9] and general particle identification [10]. Tensorflow also allows computing to be offloaded to the GPU providing large increases in performance over an equivalent CPU. A toy signal will be used again for this study and no attempt will be made to do a fit with real world data.

Due to time constraints, two important simplifications on the previous approach will be employed. Firstly only signal will be considered and no background will be generated. Secondly, this study only considers the B^0. The results for the $\overline{B}^0$ will be the same if one assumes that weak phases in the amplitudes only come from CKM matrix elements. Without that assumption a more complicated treatment is needed with both CP-averaged and CP-asymmetric observables. These will not be considered in this study.

The rest of this paper is structured as follows:

- Section 2 describes the method used in detail. Any deviations from the previous work are stated.
- Section 3 talks about the impact of changing machine learning hyperparameters on the results, and then compares how the fit for each coefficient performed.

- Section 4 discusses uncertainties in the method, the validity of the results, potential for improvement, the sensitivity to NP and compares this study to previous work.
- Section 5 considers the impact of this study and summarises the next steps.

2 Method

2.1 Observables

2.1.1 P-wave

The differential decay width for the P-wave contribution to $B^0 \to K^{*0}\mu^+\mu^-$ decay is defined as [8]

$$\begin{aligned}\frac{\mathrm{d}^4\Gamma_P[B^0\to K^{*0}\mu^+\mu^-]}{\mathrm{d}\cos\theta_\ell\,\mathrm{d}\cos\theta_K\,\mathrm{d}\phi\,\mathrm{d}q^2} =& \frac{9}{32\pi}[\, J_{1s}\sin^2\theta_K + J_{1c}\cos^2\theta_K + J_{2s}\sin^2\theta_K\cos 2\theta_\ell + \\ & J_{2c}\cos^2\theta_K\cos 2\theta_\ell + J_3\sin^2\theta_K\sin^2\theta_\ell\cos 2\phi + \\ & J_4\sin 2\theta_K\sin 2\theta_\ell\cos\phi + J_5\sin 2\theta_K\sin\theta_\ell\cos\phi + \\ & J_{6s}\sin^2\theta_K\cos\theta_\ell + J_7\sin 2\theta_K\sin\theta_\ell\sin\phi + \\ & J_8\sin 2\theta_K\sin 2\theta_\ell\sin\phi + J_9\sin^2\theta_K\sin^2\theta_\ell\sin 2\phi\,]\end{aligned} \tag{2}$$

The definitions of the angles θ_K, θ_ℓ and ϕ can be found in [11]. The J_i terms are the angular observables and are defined as

$$\begin{aligned}J_{1s} &= \frac{(2+\beta_\mu^2)}{4}\left[|A_\perp^L|^2 + |A_\parallel^L|^2 + (L\to R)\right] + \frac{4m_\mu^2}{q^2}\mathrm{Re}(A_\perp^L A_\perp^{R*} + A_\parallel^L A_\parallel^{R*}) \\ J_{1c} &= |A_0^L|^2 + |A_0^R|^2 + \frac{4m_\mu^2}{q^2}\left[2\mathrm{Re}(A_0^L A_0^{R*})\right] \\ J_{2s} &= \frac{\beta_\mu^2}{4}\left[|A_\perp^L|^2 + |A_\parallel^L|^2 + (L\to R)\right] \\ J_{2c} &= -\beta_\mu^2\left[|A_0^L|^2 + (L\to R)\right] \\ J_3 &= \frac{\beta_\mu^2}{2}\left[|A_\perp^L|^2 - |A_\parallel^L|^2 + (L\to R)\right] \\ J_4 &= \frac{\beta_\mu^2}{\sqrt{2}}\left[\mathrm{Re}(A_0^L A_\parallel^{L*}) + (L\to R)\right] \\ J_5 &= \sqrt{2}\beta_\mu\left[\mathrm{Re}(A_0^L A_\perp^{L*}) - (L\to R)\right]\end{aligned} \tag{3}$$

$$
\begin{aligned}
J_{6s} &= 2\beta_\mu \Big[\mathrm{Re}(A_\parallel^L A_\perp^{L*}) - (L \to R) \Big] \\
J_7 &= \sqrt{2}\beta_\mu \Big[\mathrm{Im}(A_0^L A_\parallel^{L*}) - (L \to R) \Big] \\
J_8 &= \frac{\beta_\mu^2}{\sqrt{2}} \Big[\mathrm{Im}(A_0^L A_\perp^{L*}) + (L \to R) \Big] \\
J_9 &= \beta_\mu^2 \Big[\mathrm{Im}(A_\parallel^{L*} A_\perp^L) + (L \to R) \Big]
\end{aligned}
$$

with $\beta_\mu^2 = (1 - 4m_\mu^2/q^2)$. The $(L \to R)$ terms are the same as the LHS but with *right* amplitudes instead of *left*. The $A_{\parallel,\perp,0}^{L,R}$ terms are complex amplitudes and are functions of the dimuon invariant mass squared, q^2. The $\parallel$, $\perp$ and 0 superscripts correspond to the polarisation of the $\overline{K}^{*0}$. The L and R subscripts correspond to the chirality of the dimuon system. A_t has been ignored which is valid in the massless limit [6] ($q^2 \gg 4m_\mu^2$) which is suitable for $q^2 > 1\,\mathrm{GeV}^2/c^4$ [3]. β_μ and m_μ terms, however were left in the observables. Scalar contributions have been neglected, which removed the A_S amplitude [6].

In the massless limit, the basis set of P_i observables with reduced reliance on form factors is given by [6]

$$
O_{m_\mu=0} = \left\{ \frac{d\Gamma}{dq^2}, A_{\mathrm{FB}}, P_1, P_2, P_3, P_4, P_5, P_6 \right\} \tag{4}
$$

The A_{FB} term, which is the forward-backward asymmetry, and the clean P_i observables can be simply related to the J_i angular observables with the following relations:

$$
\begin{aligned}
A_{\mathrm{FB}} &= -\frac{3J_{6s}}{3J_{1c} + 6J_{1s} - J_{2c} - 2J_{2s}} \\
P_1 &= \frac{J_3}{2J_{2s}} \\
P_2 &= \beta_\mu \frac{J_{6s}}{8J_{2s}} \\
P_3 &= -\frac{J_9}{4J_{2s}} \\
P_4 &= \frac{\sqrt{2}J_4}{\sqrt{-J_{2c}(2J_{2s} - J_3)}} \\
P_5 &= \frac{\beta_\mu J_5}{\sqrt{-2J_{2c}(2J_{2s} + J_3)}} \\
P_6 &= -\frac{\beta_\mu J_7}{\sqrt{-2J_{2c}(2J_{2s} - J_3)}}
\end{aligned} \tag{5}
$$

2.1.2 S-wave

The S-wave contribution to the decay width is given by [8]

$$\begin{aligned}\frac{\mathrm{d}^4\Gamma_S}{\mathrm{d}\cos\theta_\ell\,\mathrm{d}\cos\theta_K\,\mathrm{d}\phi\,\mathrm{d}q^2} =& \frac{9}{32\pi}[\ J'_{1c}(1-\cos 2\theta_\ell)\ + J''_{1c}\cos\theta_K(1-\cos 2\theta_\ell) + \\ & J'_4\sin 2\theta_\ell\sin\theta_K\cos\phi\ + J'_5\sin\theta_\ell\sin\theta_K\cos\phi + \\ & J'_7\sin\theta_\ell\sin\theta_K\sin\phi\ + J'_8\sin 2\theta_\ell\sin\theta_K\sin\phi\ ]\end{aligned} \tag{6}$$

with the J'_i observables given by

$$\begin{aligned}J'_{1c} &= \frac{1}{3}|A^L_{00}|^2 + \frac{1}{3}|A^R_{00}|^2 \\ J''_{1c} &= \frac{2}{\sqrt{3}}\Big[\mathrm{Re}(A^L_{00}A^{L*}_0) + (L\to R)\Big] \\ J'_4 &= \sqrt{\frac{2}{3}}\Big[\mathrm{Re}(A^L_{00}A^{L*}_\parallel) + (L\to R)\Big] \\ J'_5 &= 2\sqrt{\frac{2}{3}}\Big[\mathrm{Re}(A^L_{00}A^{L*}_\perp) - (L\to R)\Big] \\ J'_7 &= 2\sqrt{\frac{2}{3}}\Big[\mathrm{Im}(A^L_{00}A^{L*}_\parallel) - (L\to R)\Big] \\ J'_8 &= \sqrt{\frac{2}{3}}\Big[\mathrm{Im}(A^L_{00}A^{L*}_\perp) + (L\to R)\Big]\end{aligned} \tag{7}$$

In order to verify that the equations for P-wave and S-wave observables had been correctly implemented, two different equations were used for calculating the S-wave fraction. The first relied on the ratio of amplitudes as used in [8]

$$F_{\mathrm{S}}(q^2) = \frac{|A^L_{00}|^2 + |A^R_{00}|^2}{|A^L_\parallel|^2 + |A^L_\perp|^2 + |A^L_0|^2 + |A^L_{00}|^2 + (L\to R)} \tag{8}$$

and the second method used the ratio of angle-integrated decay widths and was defined by

$$F_{\mathrm{S}}(q^2) = \frac{\dfrac{\mathrm{d}^4\Gamma_S}{\mathrm{d}q^2}}{\dfrac{\mathrm{d}^4\Gamma_P}{\mathrm{d}q^2} + \dfrac{\mathrm{d}^4\Gamma_S}{\mathrm{d}q^2}} \tag{9}$$

Apart from negligible differences caused by approximations used, these two methods produced the same result which can be seen on the left of Fig. 4.

2.2 Anzatz

Both the real and imaginary parts of all the A_i spin amplitudes were approximated by the anzatz

$$\alpha + \beta q^2 + \frac{\gamma}{q^2} \tag{10}$$

where the α, β and γ parameters are real numbers. This anzatz provides a good fit to P-wave amplitudes generated by flavio and was used in the previous work as it provided a good fit to data generated by the EOS program [8].

2.3 Mass dependence

The P-wave contribution was attributed to the $K^{*0}(892)$ resonance. For simplicity the S-wave contribution near the $K^{*0}(892)$ mass was only considered to come from $K^{*0}(700)$. This was in line with the previous work [8] (note that the resonance is referred to as $\kappa(600)$ in that paper). In both cases the K^{*0} undergoes the decay $K^{*0} \to K^+\pi^-$. To estimate the S-wave fraction, each amplitude product in the angular observables was modified with a relativistic Breit-Wigner distribution for the $K\pi$ system

$$A_i A_j^* \to A_i A_j^* \int g_i(m_{K\pi}) g_j^*(m_{K\pi}) dm_{K\pi} \tag{11}$$

The $g_i(m_{K\pi})$ amplitudes were taken as

$$g_i(m_{K\pi}) = \frac{1}{(m_{K^{*0}}^2 - m_{K\pi}^2) - i m_{K^{*0}} \Gamma(m_{K\pi})} \tag{12}$$

where $m_{K^{*0}}$ is the mass of the parent $K^{*0}(700)$ or $K^{*0}(892)$ and $\Gamma(m_{K\pi})$ is the mass-dependent decay width defined as

$$\Gamma(m_{K\pi}) = \Gamma_0 \left(\frac{q}{q_0}\right)^{2L+1} \left(\frac{m_{K^{*0}}}{m_{K\pi}}\right) \chi^2(q, r) \tag{13}$$

where Γ_0 is the energy independent decay width, q is the momentum of daughter particles in the rest frame of the K^{*0}, q_0 is q at the mass of the K^{*0}, L is the angular momentum and χ is the barrier factor. For $L = 0$ (S-wave) the barrier factor was set as 1, and for $L = 1$ (P-wave) defined as

$$\chi = \sqrt{\frac{1 + z_0^2}{1 + z^2}} \tag{14}$$

with $z = |q|r$ and r, the radius of a hadron, defined as 4 GeV^{-1}. Values used for both the P-wave and S-wave contributions can be found in Table 1.

The $m_{K\pi}$ integration in Eq. 11 was calculated between $\pm$ 100 MeV of $m_{K^{*0}(892)}$ and the amplitudes were normalised to the value found by integrating from $m_{K^+} + m_{\pi^-}$ to infinity. The resultant values for the Breit-Wigner modifications can be found in Table 2.

Contribution	K^{*0} Resonance	$m_{K^{*0}}$	Γ_0
P-wave	892	895.55 MeV	47.3 MeV
S-wave	700	824 MeV	478 MeV

Table 1: 2018 PDG values used for the mass and decay width of P-wave and S-wave contributions [12].

Integral	Value		
	2018 PDG	2012 PDG	Previous Work
$\int \lvert g_{K^{*0}(892)}\rvert^2 dm_{K\pi}$	0.888	0.886	0.80
$\int \lvert g_{K^{*0}(700)}\rvert^2 dm_{K\pi}$	0.296	0.127	0.18
$\int g_{K^{*0}(892)} g^*_{K^{*0}(700)} dm_{K\pi}$	0.640 + 0.00968 i	0.620 - 0.0735 i	0.22 - 0.23 i

Table 2: Values of the integrated relativistic Breit-Wigner distributions used to modify the J_i angular observables that were produced from Eq. 12 with 2018 PDG [12] data, Eq. 12 with 2012 PDG [13] data, and the values used in the previous work [8]. The 2018 PDG values were used in this study.

It was not possible to produce the same values as those used in the previous work [8] with 2012 PDG [13] masses and decay widths that were used in that study. This can been seen in Table 2. However it can be seen from Fig. 2 that Eq. 12 does produce the expected distribution shapes. Because the method used to generate the previous value was not known, it was decided that Eq. 12 with more up to date 2018 PDG [12] masses and decay widths would be used.

2.4 PDF

The total decay rate was taken as the sum of Eq. 2 and Eq. 6

$$\frac{\mathrm{d}^4\Gamma}{\mathrm{d}\cos\theta_\ell\,\mathrm{d}\cos\theta_K\,\mathrm{d}\phi\,\mathrm{d}q^2} = \frac{\mathrm{d}^4\Gamma_P}{\mathrm{d}\cos\theta_\ell\,\mathrm{d}\cos\theta_K\,\mathrm{d}\phi\,\mathrm{d}q^2} + \frac{\mathrm{d}^4\Gamma_S}{\mathrm{d}\cos\theta_\ell\,\mathrm{d}\cos\theta_K\,\mathrm{d}\phi\,\mathrm{d}q^2} \tag{15}$$

The PDF used for this decay was

$$\mathrm{PDF}(\cos\theta_\ell, \cos\theta_K, \phi, q^2) = \frac{\dfrac{\mathrm{d}^4\Gamma}{\mathrm{d}\cos\theta_\ell\,\mathrm{d}\cos\theta_K\,\mathrm{d}\phi\,\mathrm{d}q^2}}{\displaystyle\int_{q^2_{min}}^{q^2_{max}} \frac{\mathrm{d}^4\Gamma}{\mathrm{d}q^2} dq^2} \tag{16}$$

where the denominator was chosen for normalisation. The integrand within it was found by analytically integrating Eq. 15 over the angles and is given by

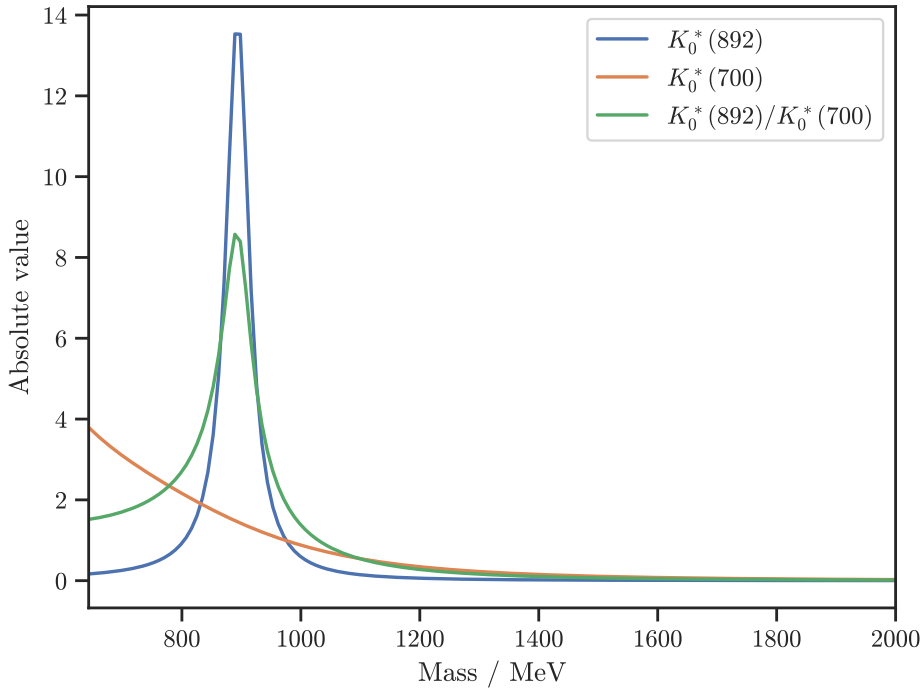

Figure 2: Breit-Wigner distributions showing the absolute value of the 2018 PDG integrands in Table 2. The blue curve corresponds to P-wave integrand, orange is the S-wave integrand, and the green curve is the integrand for mixed terms. The lower mass limit is the combined rest mass of K^+ and π^-.

$$\frac{\mathrm{d}^4\Gamma}{\mathrm{d}q^2} = \frac{1}{4}(6J_{1s} + 3J_{1c} - 2J_{2s} - J_{2c}) + 3J'_{1c} \tag{17}$$

It is implicit that the PDF is ultimately dependent on the particular anzatz parameters used. As the integral in the PDF needed recomputing for each change in coefficients, it was performed numerically using a trapezoid rule method for performance reasons. The step size of this was tested for a range of coefficient inputs to ensure that the result was no more than 0.1% different from the value found from a Dormand-Prince method.

The q^2_{min} value was set as $1\,\mathrm{GeV}^2/c^4$ as per the previous work [8] to avoid light resonances below this. The q^2_{max} value was set as $8\,\mathrm{GeV}^2/c^4$, higher than the $6\,\mathrm{GeV}^2/c^4$ used in previous work, but this was considered low enough to avoid $c\bar{c}$ resonances above this. It was hoped this increase in q^2 range would help the fit converge better.

2.5 Basis fixing

The angular distribution is invariant under a continuous U(2) rotation of the amplitudes involving 4 angular variables [14]. The transformation is as follows

$$n_i' = \begin{pmatrix} e^{i\phi_L} & 0 \\ 0 & e^{-i\phi_R} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cosh i\omega & -\sinh i\omega \\ -\sinh i\omega & \cosh i\omega \end{pmatrix} n_i \tag{18}$$

with the basis vectors n_i as

$$n_\parallel = \begin{pmatrix} A_\parallel^L \\ A_\parallel^{R*} \end{pmatrix}\ ,\ n_\perp = \begin{pmatrix} A_\perp^L \\ -A_\perp^{R*} \end{pmatrix}\ ,\ n_0 = \begin{pmatrix} A_0^L \\ A_0^{R*} \end{pmatrix}, \tag{19}$$

with the angles ϕ_L, ϕ_R, θ and ω defined as

$$\begin{aligned}
\tan 2\omega &= 2\frac{\mathrm{Im}(A_0^R)\mathrm{Re}(A_0^L) + (L \leftrightarrow R)}{|A_0^R|^2 - |A_0^L|^2} \\
\tan\theta &= \frac{\mathrm{Re}(A_0^R) + \mathrm{Im}(A_0^L)\tan\omega}{-\mathrm{Re}(A_0^L) + \mathrm{Im}(A_0^R)\tan\omega} \\
\tan\phi_L &= \frac{\mathrm{Im}(A_0^L) + \mathrm{Im}(A_0^R)\tan\theta - [\mathrm{Re}(A_0^R) - \mathrm{Re}(A_0^L)\tan\theta]\tan\omega}{-\mathrm{Re}(A_0^L) + \mathrm{Re}(A_0^R)\tan\theta + [\mathrm{Im}(A_0^R) + \mathrm{Im}(A_0^L)\tan\theta]\tan\omega} \\
\tan\phi_R &= \frac{\mathrm{Im}(A_\perp^R) + \mathrm{Im}(A_\perp^L)\tan\theta - [\mathrm{Re}(A_\perp^L) - \mathrm{Re}(A_\perp^R)\tan\theta]\tan\omega}{-\mathrm{Re}(A_\perp^R) + \mathrm{Re}(A_\perp^L)\tan\theta + [\mathrm{Im}(A_\perp^L) + \mathrm{Im}(A_\perp^R)\tan\theta]\tan\omega},
\end{aligned} \tag{20}$$

The angles introduce degeneracy to the angular distribution and so to produce a unique solution, the basis needs be fixed by setting 4 variables to a fixed value. For this study the following was chosen

$$\mathrm{Re}(A_0^R) = \mathrm{Im}(A_0^R) = \mathrm{Im}(A_0^L) = \mathrm{Im}(A_\perp^R) = 0 \tag{21}$$

This choice was motivated by a desire to avoid a discontinuity with the $\mathrm{Im}(A_0^L)$ component around $2\,\mathrm{GeV}^2/c^4$ as discussed in the previous work [8]. The transformation was performed prior to inputting signal coefficients to the code, and all fits were done using this fixed basis.

2.6 Signal generation

In this paper two different physics models were used for the signal generation coefficients: the standard model (SM) and a new physics (NP) model with modified Wilson coefficients of $\mathcal{C}_9 = -1.027$ and $\mathcal{C}_{10} = 0.498$. The choice of Wilson coefficients for the NP model was based on fits to data and is similar to best fit values recently published after Moriond 2019 [1]. The list of coefficients used can be found in Table 3.

For $A_{\parallel,\perp,0}^{L,R}$ the three anzatz parameters for each complex component were found through a fit using flavio [15]. It was not possible to find the $A_{00}^{L,R}$ values using this method so they were approximated in both sets by the constant complex value $1 + 1i$. The previous work [8] also approximated the $A_{00}^{L,R}$ values as a constant, but as it was not stated what value was used, this placeholder was chosen as it gives the correct order for the S-wave fraction of ($\mathcal{O}(10\%)$) as can be seen in the left side of Fig. 4.

Component	Model					
	SM			NP		
	α	β	γ	α	β	γ
$\mathrm{Re}(A_{\parallel}^{L})$	-4.17810	-0.15184	+6.81832	-3.42775	-0.12410	+6.04528
$\mathrm{Im}(A_{\parallel}^{L})$	+0.00859	-0.00182	+0.46607	+0.00934	-0.00199	+0.50341
$\mathrm{Re}(A_{\parallel}^{R})$	-0.23538	-0.00432	+8.00375	-0.25087	-0.00518	+8.63675
$\mathrm{Im}(A_{\parallel}^{R})$	+0.16564	-0.01310	-0.30668	+0.22209	-0.01742	-0.52807
$\mathrm{Re}(A_{\perp}^{L})$	+3.88641	+0.08527	-8.19745	+3.06464	+0.07852	-8.84114
$\mathrm{Im}(A_{\perp}^{L})$	-0.09505	+0.00793	-0.07297	-0.11366	+0.00929	-0.04762
$\mathrm{Re}(A_{\perp}^{R})$	-0.42358	+0.02730	-7.14745	-0.93327	+0.01687	-6.31856
$\mathrm{Im}(A_{\perp}^{R})$	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000
$\mathrm{Re}(A_{0}^{L})$	+7.20276	-0.22782	+9.89863	+5.88288	-0.18442	+8.10140
$\mathrm{Im}(A_{0}^{L})$	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000
$\mathrm{Re}(A_{0}^{R})$	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000
$\mathrm{Im}(A_{0}^{R})$	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000	+0.00000
$\mathrm{Re}(A_{00}^{L})$	+1.00000	+0.00000	+0.00000	+1.00000	+0.00000	+0.00000
$\mathrm{Im}(A_{00}^{L})$	+1.00000	+0.00000	+0.00000	+1.00000	+0.00000	+0.00000
$\mathrm{Re}(A_{00}^{R})$	+1.00000	+0.00000	+0.00000	+1.00000	+0.00000	+0.00000
$\mathrm{Im}(A_{00}^{R})$	+1.00000	+0.00000	+0.00000	+1.00000	+0.00000	+0.00000

Table 3: Coefficient values of signal anzatz parameters used for each model in this study.

Unfortunately it was found that the flavio fit method did not produce the same shaped observables as flavio itself generated. This was because flavio internally does not work with transversity amplitudes, and the attempted conversion to them did not produce the required result. An example of this can be seen in Fig. 3.

There were also differences with the S-wave fraction shape compared with the previous paper. Although the differences in Breit-Wigner values and the placeholder anzatz for $A_{00}^{L,R}$ contribute to this, the main problem is believed to originate from the generated coefficients. This can be seen in Fig. 4.

Furthermore the incorrect coefficients then led to a differently shaped angle-integrated differential decay rate versus what was expected from flavio. This is shown in Fig. 5.

For each fit to a set of signal coefficients, signal events were randomly generated using Eq. 16 with a Monte-Carlo accept-reject method. Unless otherwise stated, the event count for each fit was set as 2,400. This number was stated in the previous work as the expected yield from Run-II [8]. No increase in the event count was performed to account for the increase in the q^2 range in this work, and this was to allow better comparison of performance. An example of the signal produced in this study using SM coefficients can

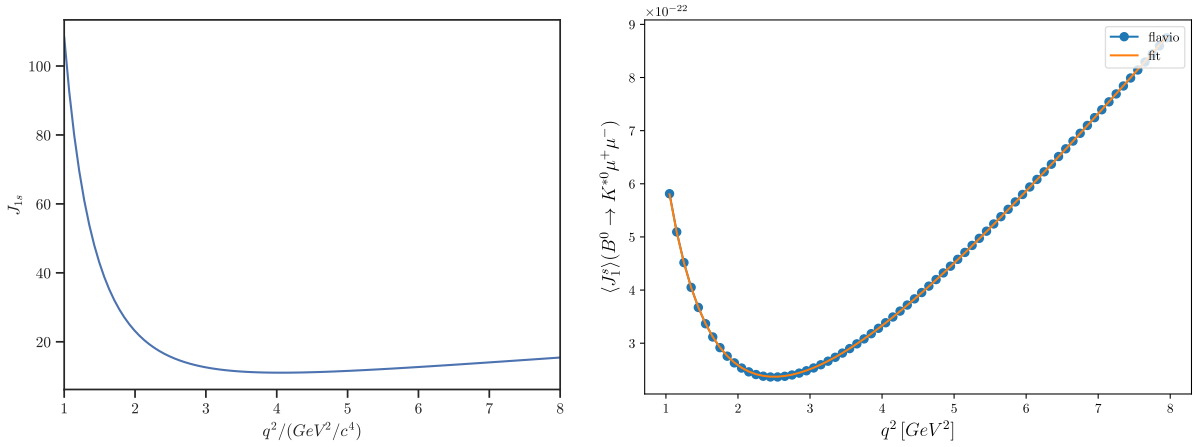

Figure 3: Comparison of the J_{1s} angular observable curve shape produced by the coefficients used in this study (left) and what comes out of flavio (right). Both were produced from the NP model. The most clear difference is the missing increase after $\sim 2.5\,\mathrm{GeV}^2/c^4$. The y-axis scale shown is irrelevant for this discussion.

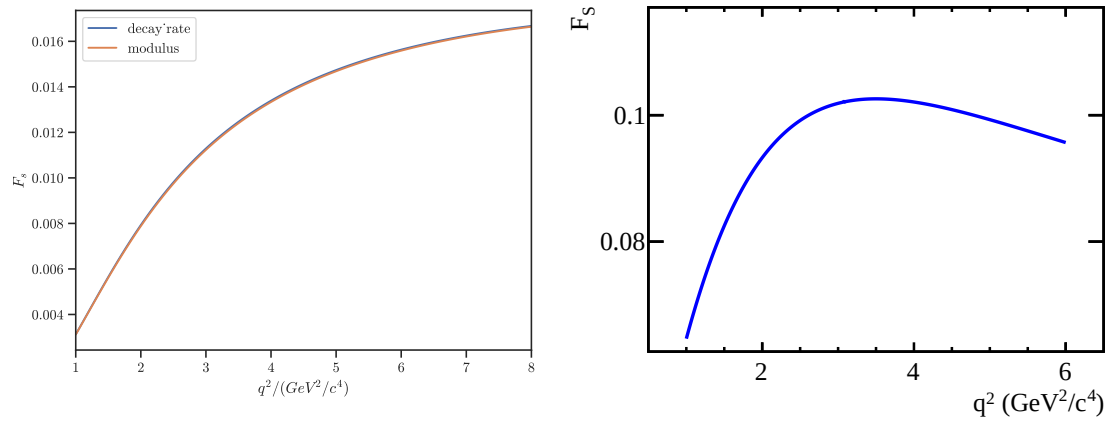

Figure 4: Comparison of the S-wave fraction to the decay from this study (left) and the previous study [8] (right). Both were produced from the SM model. Although the order of magnitude is correct, the shape doesn't show a maxima in this study. The differences are primarily thought to stem from issues generating the signal coefficients. The blue line on the left corresponds to using Eq. 9 and the orange line is from Eq. 8.

be found in Fig. 6.

Approximate symmetries are expected for the low signal counts used in this study for models with no right-handed currents. These were expected to correspond to the SM signal coefficients, but not the NP ones. For the basis chosen in this paper, the following left-handed amplitude transformation would be expected to display the symmetry [8]:

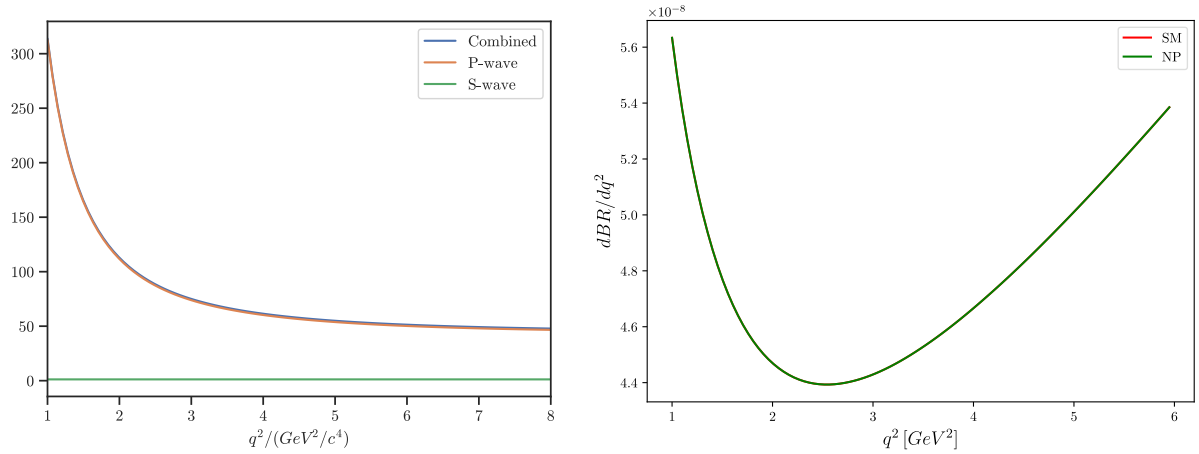

Figure 5: Comparison of the angle-integrated differential decay rate curve shape produced by the NP coefficients used in this study (left) and what comes out of flavio (right). The left plot contains the P and S-wave contributions separately, as well as the combined contribution that is shown on the right. There is clearly a missing minima in the decay rate of this study. The y-axis scale shown is irrelevant for this discussion.

$$\begin{aligned} A_{\parallel}^{L} &\to -A_{\perp}^{L} \\ A_{\perp}^{L} &\to -\frac{A_{\parallel}^{L}}{2} \end{aligned} \tag{22}$$

To test for these symmetries, plots of the signal were generated again for both models with the transformation applied. These can be seen for $\cos\theta_\ell$ in the NP model in Fig. 7. The approximate symmetry is seen in the NP model which is unexpected. They also exist in the SM model. This suggests that both signal models used in this study correspond to models with zero of negligible right-handed currents.

Another approximate symmetry exists through a right-handed transformation of:

$$A_{\parallel}^{R} \leftrightarrow -A_{\perp}^{R} \tag{23}$$

This again would only be expected to be seen for models with no right-handed amplitudes but is seen in both. This symmetry will appear in the confidence plots discussed in Section 4.1.1.

2.7 Fitting algorithm

To perform a fit, Tensorflow was made to find the coefficients that minimised the normalised negative log likelihood which was defined as

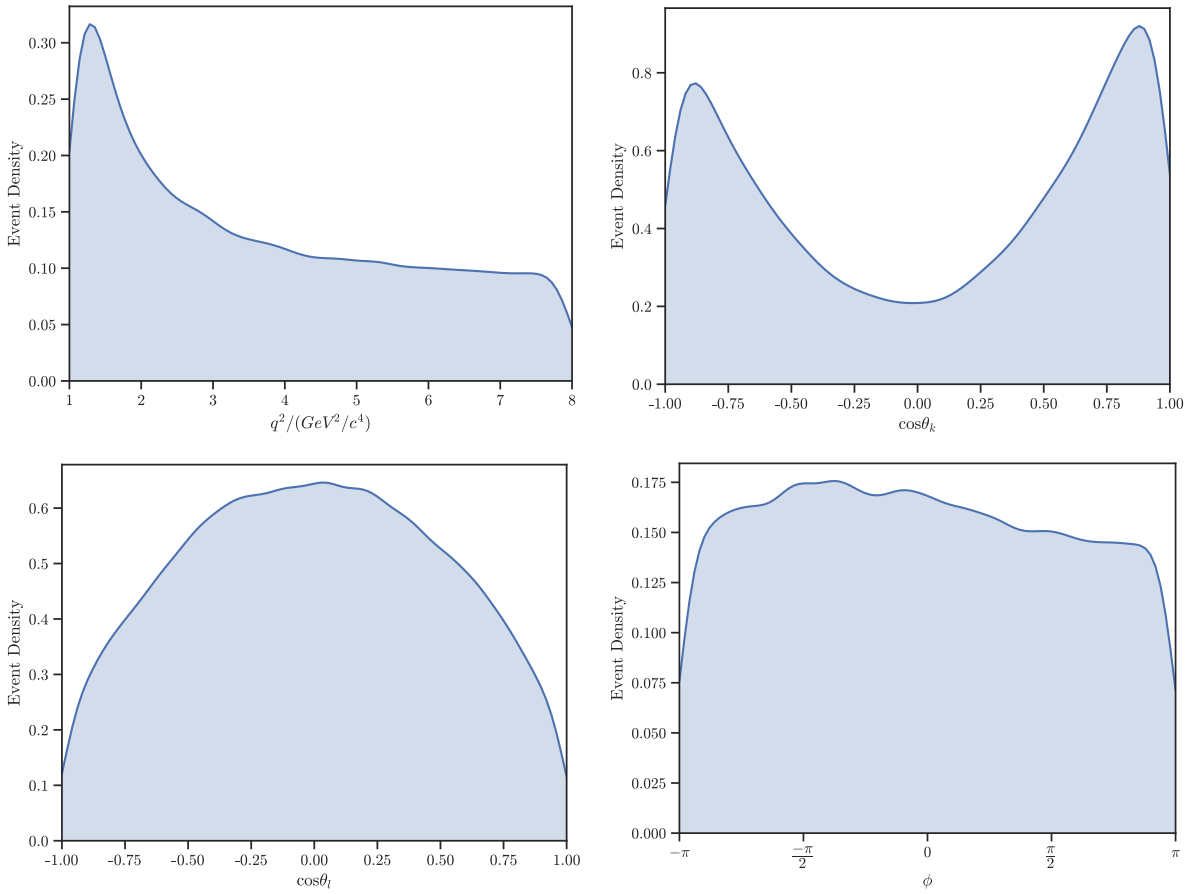

Figure 6: Density of SM signal events shown for q^2 (top left), $\cos\theta_K$ (top right) $\cos\theta_\ell$ (bottom left) and ϕ (bottom right). 100,000 events were generated to produce these plots in order to smooth the shapes.

$$- \frac{\sum_i \log(\text{PDF}(\cos\theta_{\ell,i}, \cos\theta_{K,i}, \phi_i, q_i^2))}{N_{events}} \tag{24}$$

where the i subscript represents the independent variables for each event in this fit, and most often, $N_{events} = 2,400$. Again it is implicit that the PDF depends on the particular values of the anzatz parameters for this fit iteration. The normalisation was added so that the magnitude and thus optimiser tuning would be unaffected if N_{events} was changed.

The optimisation process was iterated until the maximum gradient of any anzatz coefficient was 5×10^{-7}, or until 20,000 steps was achieved, in which case the fit was restarted with the same signal events but with the fit coefficient values reinitialised.

Three different algorithms were used for initialising the coefficients on each fit. These were:

- TWICE_LARGEST_SIGNAL_SAME_SIGN: Coefficients initialised from 0 to 2× the largest signal coefficient value for each coefficient in either signal model

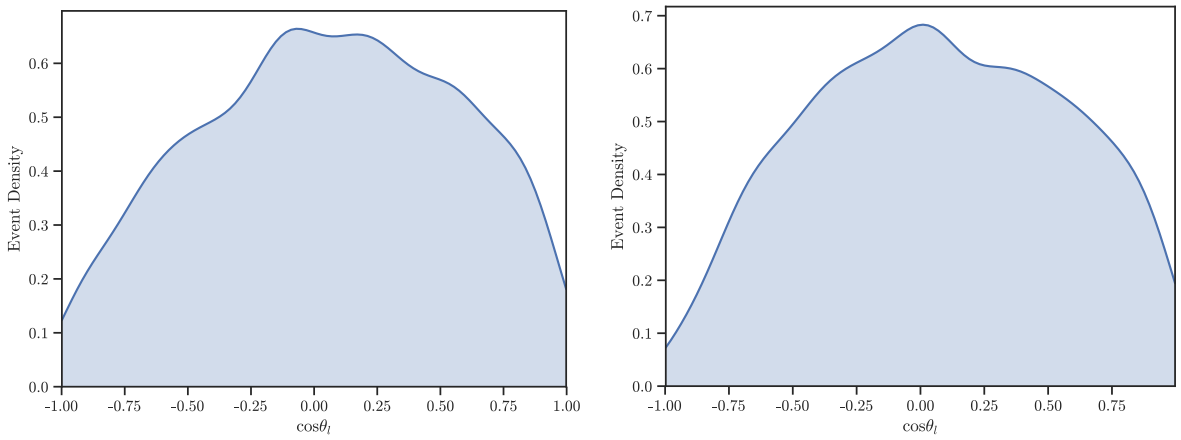

Figure 7: NP signal plots for $\cos\theta_\ell$ with no transformation (left) and with the transformation of Eq.22 (right). A event count of 2,400 was used. Both show a similar shape suggesting the NP model used in this study mistakenly corresponds to one with zero or negligible right-handed currents.

- TWICE_CURRENT_SIGNAL_ANY_SIGN: Coefficients initialised from $-2\times$ to $+2\times$ the signal coefficient values used for this fit
- CURRENT_SIGNAL: Coefficients initialised to the signal coefficient values

The TWICE_CURRENT_SIGNAL_ANY_SIGN algorithm was implemented to demonstrate that the code showed up the discrete symmetries in the angular distribution as can be seen in Fig. 8. These arise because the angular observables are produced from products of amplitudes, therefore the transformation $A_i \to -A_i$ for all i can be applied whilst still keeping the observables the same.

Unless otherwise stated when discussing results, it should be assumed that the TWICE_LARGEST_SIGNAL_SAME_SIGN algorithm has been used. This was implemented as a trade-off to demonstrate that fit results were not dependent on a prior knowledge of the particular signal model used, but applying some constraints to ensure coefficients close to 0 in both models were not started off at high values where the fit would struggle to converge. This algorithm mainly avoids the discrete symmetries except for coefficients close to 0, and this was deliberately done to simplify the analysis of uncertainties later on.

Only the P-wave coefficients for unfixed amplitudes in this basis as well as the A_{00} α parameters were treated as free parameters for fitting. This resulted in 28 coefficients that needed to be determined for each fit (24 for P-wave and 4 for S-wave).

2.8 Hardware & software

All performance numbers given later were based on code running on an Intel i7-7700HQ 2.80 GHz CPU and a NVIDIA Quadro M1200 Mobile GPU. Tensorflow v1.14 was used,

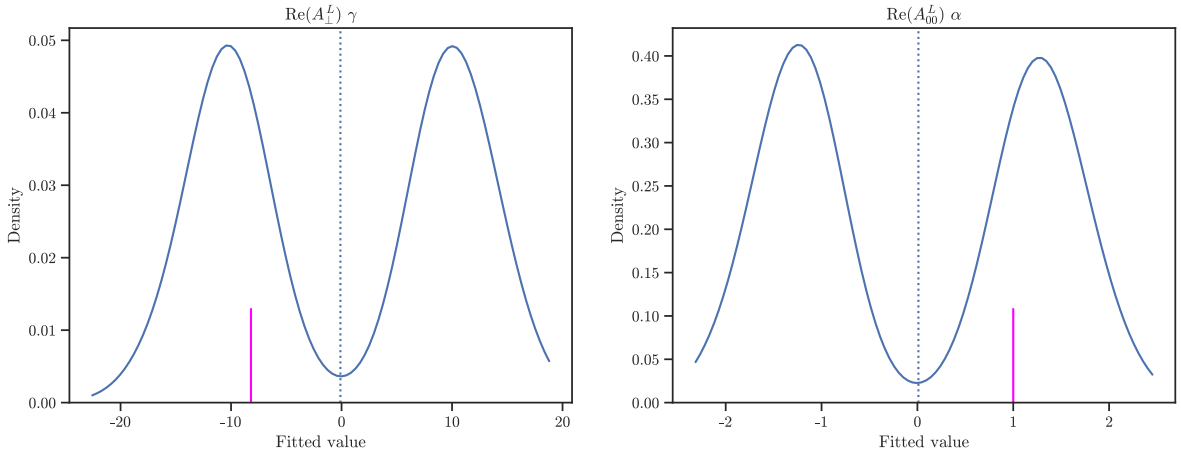

Figure 8: Distribution of results from an ensemble of fits for the Re($A_\perp^L$) γ (left) and Re(A_{00}^L) α (right) coefficients when using the TWICE_CURRENT_SIGNAL_ANY_SIGN algorithm and the SM model. The discrete symmetry discussed in the text is visible in the double peaks. The magenta line signifies the signal value that was used in the fits.

but with Tensorflow v2 behaviour compatibility enabled. The code for this project has been published on GitHub under an MIT license and can be found at [16].

3 Results

3.1 Optimiser algorithm

Various optimisation algorithms were tried in this study. It was not possible to get fits to converge quickly with either the SGD [17], RMSProp [18] or Nadam [19] algorithms. Adam [20] could sometimes quickly converge, but often suffered from a problem of "exploding gradients". This was where the coefficients that converged to values earliest would eventually start diverging again before all coefficients had converged. The reason for this was a divide by an increasingly small number in the Adam algorithm. Changing the parameter ϵ which is meant to provide numerical stability helped with this issue, but it was still a major problem. In the end the AMSGrad [20, 21] variant of Adam was chosen as it demonstrated the ability to quickly converge without the exploding gradients.

The issue with the simpler algorithms like SGD is believed to be because the signal coefficients have orders of magnitude differences between them. A possible solution for this would be to normalise the coefficients to the same order for the algorithm, and then re-scale them for inputting into equations. An effort was made to do this early in the work but was aborted due to time constraints when the AMSGrad algorithm was found to work. Algorithms like Adam and AMSGrad implement "momentum" which will be discussed in the next subsection. This has the unfortunate effect of adding bias to the results, so revisiting the normalisation could be of interest to follow-up work.

3.2 Hyperparameters

Some of the tunable hyperparameters passed to the optimiser were found to have an effect on the fitting quality and performance. The hyperparameters that were tested in this study were the learning rate, β_1, β_2 and ϵ. The learning rate is a general hyperparameter that affects the initial step size before decay, and the other three are specific to the AMSGrad algorithm. The algorithm for AMSGrad as well as the definitions of these hyperparameters can be found in [21].

The largest effect was seen from changing the learning rate. Four rates were compared: 0.05, 0.10, 0.15 and 0.20. β_1, β_2 and ϵ were left at their default values for this comparison. The 0.05 rate was much slower to complete each NP fit at around 36.3 seconds compared to 21.4 seconds for 0.10, 17.0 seconds for 0.15 and 14.9 seconds for 0.20. With the NP model, 0.05 generally had the closest fit mean to the signal value across all coefficients, with the drift increasing with each increase in learning rate. The SM fits were more mixed with sometimes 0.10 providing the closest fit mean, and sometimes 0.05, but always with 0.15 and 0.20 last in that order. An example comparison of learning rates for $\mathrm{Re}(A_{\parallel}^{L})$ α in the the NP model can be seen in Fig. 9. It can also be seen in this figure that changing the learning rate not only changed the mean of the distribution, but also the shape. Some coefficients always had symmetrical distributions regardless of learning rate, some showed skew only for lower learning rates, and the others showed skew only for higher learning rates. There was no discernible pattern found with what coefficients showed what behaviour.

The reason the learning rate affects the skew as well as the mean is thought to be due to the "momentum" that is implemented in the algorithm. This is a method to ensure that the optimiser keeps moving towards the minimum when working with a variable without a smooth likelihood profile. Without momentum, the optimiser would have a tendency to decrease the gradients too quickly. Momentum however has the effect that if the optimum value for a coefficient is overshot, the optimiser is prevented from quickly reversing direction. If the gradients decrease too much then the optimiser either won't have time to turn around before convergence is declared, or the gradient step sizes will be very small when it does.

With the β_1 hyperparameter three values were compared: 0.85, 0.90 (default) and 0.95. The learning rate was set at 0.20 for this comparison and β_2 and ϵ were left at their default values. Across both models the 0.85 fit mean was marginally closer to the signal than 0.90, with 0.95 the worst. Additionally 0.95 seemed to also increase the fit distribution width. The default was fastest at 14.9 seconds per NP fit, compared to 18.5 seconds for 0.95 and 23.5 seconds for 0.85. A good example of the fitting difference can be seen in Fig. 10. The distribution shapes for 0.85 and 0.90 were fairly Gaussian shaped, however 0.95 often produced uneven curves.

Changing the β_2 hyperparameter showed no clear change on the fitting quality in either model when using a learning rate of 0.20 and keeping β_1 and ϵ at their default values. It did however change the fitting speed with the default 0.999 the fastest at 14.9 seconds, followed by 15.8 seconds for 0.9995 and 20.2 seconds for 0.995.

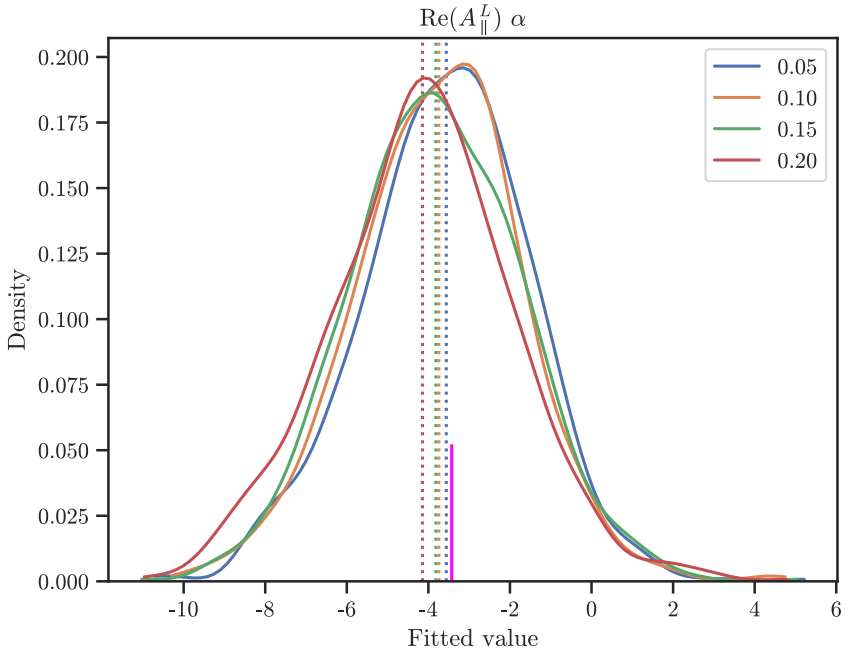

Figure 9: Comparison of the effect of changing the learning rate on an ensemble of 1000 fits for Re($A_{\parallel}^{L}$) α in the the NP model. The dotted lines represent the mean of the fit distributions and the magenta line is the signal value. β_1, β_2 and ϵ have been left at their default values. The lower values for the learning rate give the closest mean fit to the signal value. Note that the 0.15 and 0.20 distributions are fairly symmetrical however the 0.05 and 0.10 distributions have a longer left tail.

Increasing the ϵ hyperparameter from the default showed an small increase in the closeness of the fit mean and signal for both models in coefficients that had a signal value far from 0 (see left of Fig. 11). In those cases 10^{-3} performed best followed by 10^{-5} with the default of 10^{-8} last. For coefficients close to 0, 10^{-5} was generally closest with the other two values either side of it (see right of Fig. 11). The default of 10^{-8} provided the fastest fit time of 14.8 seconds, followed by 10^{-5} just behind at 16.0 seconds and with 10^{-3} much slower at 28.1 seconds. In these comparisons the learning rate was again set at 0.20 with β_1 and β_2 left at their default values. Changing ϵ also had unpredictable effects on the distribution skew.

3.3 Means, standard errors & pulls

An ensemble of 1000 fits was performed with a SM signal and a NP one. A learning rate of 0.10 was used for this data with β_1, β_2 and ϵ left at their default values. The choice of these hyperparameters was motivated by a desire to improve the fitting performance whilst still being able to generate a sufficient dataset in the project time available. For each coefficient, the mean, standard error and pull mean was calculated. All the values

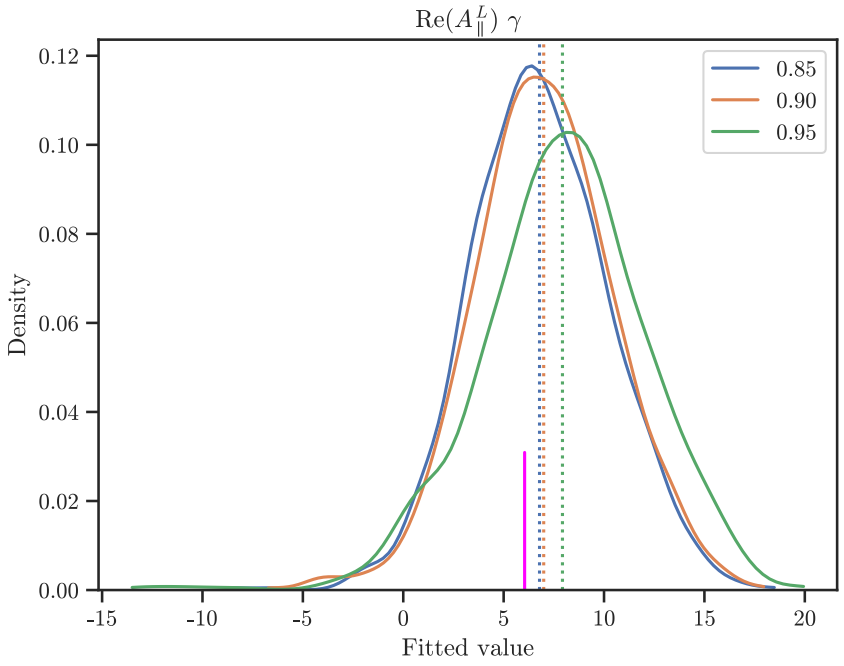

Figure 10: Comparison of the effect of changing β_1 on an ensemble of 1000 fits for Re($A_{\parallel}^{L}$) γ in the the NP model. The dotted lines represent the mean of the fit distributions and the magenta line is the signal value. The learning rate was set at 0.20, and β_2 and ϵ have been left at their default values. 0.85 is marginally better than 0.90, with 0.95 noticeably worse. Note that the 0.95 distribution is also noticeably wider than the other two.

calculated are listed in Appendix A.

The means for the SM ensemble are fairly close their signal values for most coefficients. The full table of results for this ensemble can be found in Appendix A.1. In absolute terms the coefficients that were furthest away from the signal are Re($A_{\parallel}^{L}$) γ, Re($A_{\parallel}^{R}$) α, Re(A_{0}^{L}) α and Re(A_{0}^{L}) γ which are all 0.5 to 0.6 from the signal. There is no pattern with these coefficients regarding under-estimation versus over-estimation. With the exception of Re($A_{\parallel}^{R}$) α these coefficients all have large signal values. The reason for Re($A_{\parallel}^{R}$) α having a larger difference appears to be the discrete symmetries previously discussed that lead to a fit solution with the opposite sign. As one might expect, the coefficients with the large relative differences correspond to ones with a signal value close to 0. The most extreme is Re($A_{\parallel}^{R}$) β that has a fit mean of -0.10475 versus a -0.00432 signal value - a 23.3 factor difference. A truncated table of results showing the 4 coefficients with the largest differences can be found in Table 4.

The four smallest differences for the SM ensemble can be found in Table 5. Re($A_{\perp}^{R}$) γ is particularly noteworthy as although it has a large signal value of -7.14745, the fitted value of -7.14570 is extremely close.

The NP fits were not as close to the signal values as the SM fits. The full result table can be found in Appendix A.2. They show more extreme absolute value differences.

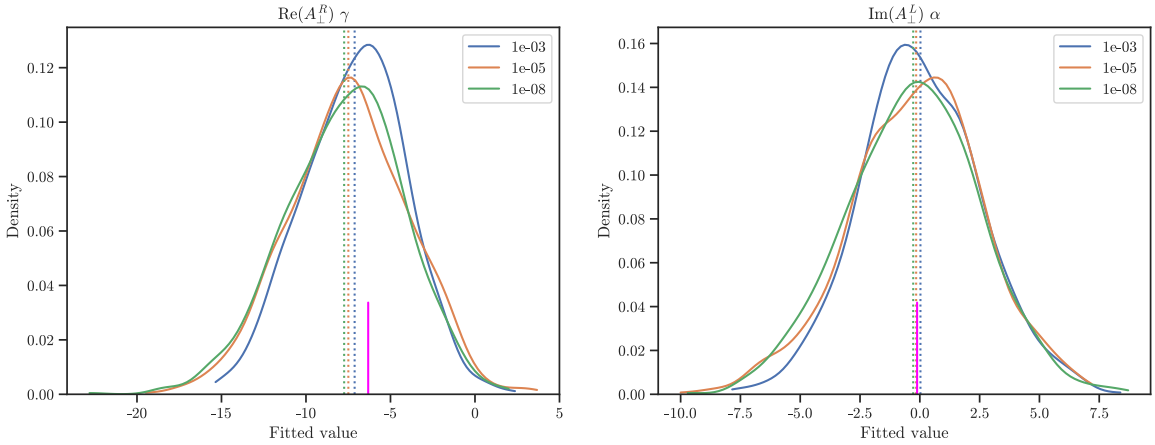

Figure 11: Comparison of the effect of changing ϵ on an ensemble of 1000 fits for Re($A_\perp^R$) γ (left) and Im($A_\perp^L$) α (right) in the the NP model. The dotted lines represent the mean of the fit distributions and the magenta line is the signal value. The learning rate was set at 0.20, and β_1 and β_2 have been left at their default values. The left distribution shows a coefficient with a signal value far from 0 where larger values of epsilon get a mean closer to the signal value. The right plot shows a coefficient with a value close to 0, in this case 10^{-5} provides the closest result with the other two values either side.

Coefficient	Signal	Mean	Difference
Re(A_0^L) γ	+9.89863	+10.48869	+0.59006
Re($A_\parallel^R$) α	-0.23538	+0.33127	+0.56665
Re(A_0^L) α	+7.20276	+7.72792	+0.52516
Re($A_\parallel^L$) γ	+6.81832	+6.29739	-0.52094

Table 4: Four largest differences between the signal and mean fit values in the SM ensemble

Coefficient	Signal	Mean	Difference
Im($A_\parallel^L$) β	-0.00182	-0.00149	+0.00033
Re($A_\perp^R$) γ	-7.14745	-7.14570	+0.00175
Im($A_\parallel^L$) α	+0.00859	+0.00086	-0.00772
Im(A_{00}^R) α	+1.00000	+0.99124	-0.00876

Table 5: Four smallest differences between the signal and mean fit values in the SM ensemble

Re($A_\perp^L$) γ, Re(A_0^L) α and Re(A_0^L) γ all have absolute differences between 1.1 and 1.4 from their signal values. Again these coefficients have large signal values in this model. It is interesting to note that Re(A_0^L) α and Re(A_0^L) γ have larger absolute differences in both models. It is true again with this model that the coefficients with the largest relative

differences correspond to signal values close to 0. Re($A_\perp^R$) γ that was so close with the SM signal considering the large value, this time is off by an absolute value of 0.78. The largest differences in this ensemble can be found in Table 6. Also of interest is that fact that the largest differences in both models correspond to real components. β coefficients don't feature in the largest differences in either ensemble, and this is believed to be because all the values used in the signal models were small for that parameter.

Coefficient	Signal	Mean	Difference
Re($A_\perp^L$) γ	-8.84114	-10.23768	-1.39653
Re(A_0^L) γ	+8.10140	+9.47912	+1.37772
Re(A_0^L) α	+5.88288	+6.97881	+1.09592
Re($A_\parallel^R$) γ	+8.63674	+9.44652	+0.80978

Table 6: Four largest differences between the signal and mean fit values in the NP ensemble

The smallest differences for the NP ensemble all correspond to β coefficients, with no large signal values appearing. This can be seen in Table 7.

Coefficient	Signal	Mean	Difference
Im($A_\parallel^R$) β	-0.01742	-0.01651	+0.00091
Im($A_\perp^L$) β	+0.00929	+0.01074	+0.00145
Im($A_\parallel^L$) β	-0.00199	+0.00349	+0.00548
Re($A_\parallel^L$) β	-0.12410	-0.13436	-0.01026

Table 7: Four smallest differences between the signal and mean fit values in the NP ensemble

A scatter plot of the differences versus the absolute signal value for both ensembles can be found in Fig. 12. The left shows this relationship broken down by model. Although the coefficients with the smallest signal values have many small differences, there isn't a clear correlation otherwise. The right side shows this relationship broken down by anzatz parameter. The real β and imaginary parameters which all have low signal values dominate the low differences, whereas the real α and γ parameter show much more of a spread. There was no clear relationship when this plot was broken down by amplitude name.

Despite the differences between the mean and signal discussed, both models show fairly low values for the standard error of each coefficient. The standard error for each coefficient was calculated as

$$\sigma_{err} = \frac{\sigma_{std}}{\sqrt{N_{fits}}} \tag{25}$$

with σ_{std} the standard deviation for that coefficient and N_{fits} set as 1000.

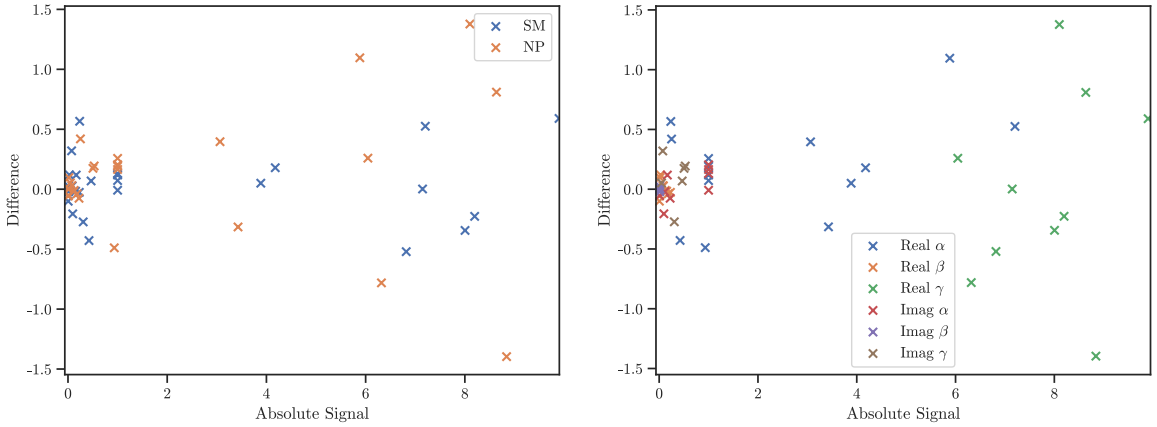

Figure 12: Plot of the difference between the fit mean and signal value versus the absolute signal value for both the SM and NP ensembles. The left side is broken down by model and the right by anzatz parameter. The real β and imaginary parameters which have low signal values are fairly consistent at having low differences, whereas the real α and γ parameter show much more of a spread.

The four highest values for the standard error in the SM ensemble can be see in Table 8. $\mathrm{Im}(A_{\parallel}^{R})$ γ shows the highest value with 0.18585. Again β coefficients are absent from the top list. Both small and large signals appear in this list. It is notable that the top three coefficients are imaginary components.

Coefficient	Signal	Mean	Difference	Std. Err
$\mathrm{Im}(A_{\parallel}^{R})$ γ	-0.30668	-0.57944	-0.27276	0.18585
$\mathrm{Im}(A_{\parallel}^{R})$ α	+0.16564	+0.28405	+0.11841	0.13308
$\mathrm{Im}(A_{\perp}^{L})$ γ	-0.07297	+0.24708	+0.32005	0.11652
$\mathrm{Re}(A_{\parallel}^{R})$ γ	+8.00375	+7.65988	-0.34386	0.10793

Table 8: Four largest standard errors in the SM ensemble

The four lowest coefficients in the SM ensemble correspond to β coefficients as can be seen in Table 9. All of the four listed are coefficients with low values for the differences and fairly low signal values.

The values of the standard errors for the NP ensemble are quite similar. The coefficients in the top four are the same ones as seen with the SM ensemble, apart from ranks 2 ($\mathrm{Im}(A_{\parallel}^{R})$ α) and 3 ($\mathrm{Im}(A_{\perp}^{L})$ γ) having the order reversed. These are listed in Table 10.

The lowest four in the NP ensemble have similar values to the SM ensemble, and have the same coefficients listed in the same order - see Table 11. The differences and signal values are also low for the lowest standard error coefficients in this ensemble.

Scatter plots of the standard error versus the absolute signal values can be found in Fig. 13. The left side shows the values broken down by model. This plot shows two

Coefficient	Signal	Mean	Difference	Std. Err
$\mathrm{Re}(A_0^L)\ \beta$	-0.22782	-0.25409	-0.02628	0.00450
$\mathrm{Re}(A_\perp^L)\ \beta$	+0.08527	+0.11211	+0.02685	0.00618
$\mathrm{Im}(A_\parallel^L)\ \beta$	-0.00182	-0.00149	+0.00033	0.00686
$\mathrm{Re}(A_\parallel^L)\ \beta$	-0.15184	-0.16856	-0.01672	0.00809

Table 9: Four smallest standard errors in the SM ensemble

Coefficient	Signal	Mean	Difference	Std. Err
$\mathrm{Im}(A_\parallel^R)\ \gamma$	-0.52807	-0.33425	+0.19382	0.17437
$\mathrm{Im}(A_\perp^L)\ \gamma$	-0.04762	+0.00466	+0.05228	0.12424
$\mathrm{Im}(A_\parallel^R)\ \alpha$	+0.22209	+0.14636	-0.07573	0.11937
$\mathrm{Re}(A_\parallel^R)\ \gamma$	+8.63674	+9.44652	+0.80978	0.10630

Table 10: Four largest standard errors in the NP ensemble

Coefficient	Signal	Mean	Difference	Std. Err
$\mathrm{Re}(A_0^L)\ \beta$	-0.18442	-0.22278	-0.03835	0.00438
$\mathrm{Re}(A_\perp^L)\ \beta$	+0.07852	+0.10878	+0.03026	0.00632
$\mathrm{Im}(A_\parallel^L)\ \beta$	-0.00199	+0.00349	+0.00548	0.00689
$\mathrm{Re}(A_\parallel^L)\ \beta$	-0.12410	-0.13436	-0.01026	0.00782

Table 11: Four smallest standard errors in the NP ensemble

distinct lines. One vertical line for coefficients that have a range of standard errors but low signal values, and another positive slant line where there is a loose correlation between the signal value and the standard error. There is also a cluster of values with low signals and low standard errors. There is no obvious difference in this plot between the two models. The right plot shows this broken down by anzatz parameter instead. From this it can be generally seen that imaginary α and γ parameters are generally the ones with a low signal value but a range of standard errors. Real α followed by real γ make up the line with loose correlation. The cluster near the origin is mainly explained by β parameters. No similar pattern was shown by breaking this plot down by amplitude name.

The pull mean was calculated to give an indication of how coefficients were biased and was an average of the pull for each individual coefficient across all fits. The individual coefficient pulls for each fit were calculated as

$$\text{Pull} = \frac{\text{Fitted Value} - \text{Signal Value}}{\sigma_{err}} \tag{26}$$

with σ_{err} the standard error. This method of calculating pulls differed from the previous

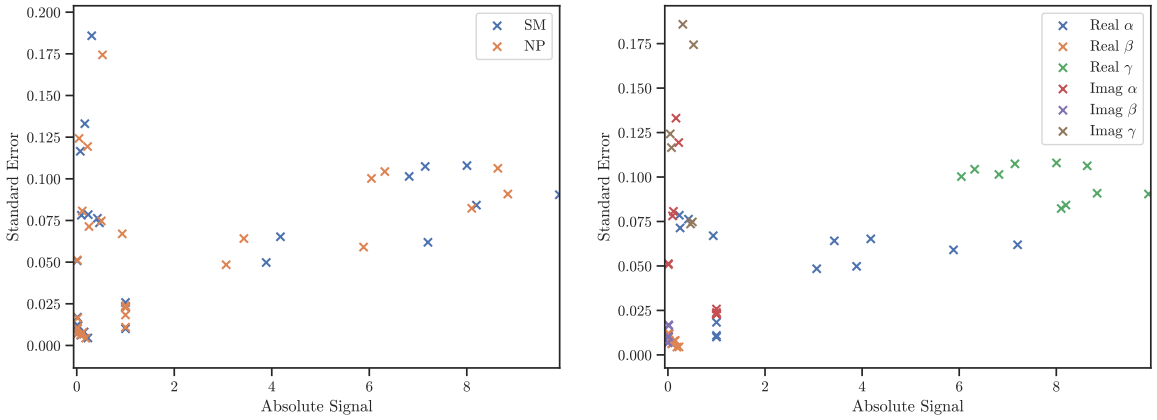

Figure 13: Plot of the standard error versus the absolute signal value. The left side shows the plot broken down by model whereas the right side shows it broken down by anzatz parameter. Coefficients with a low signal value show a range of standard errors, but coefficients with a higher signal value show a loose correlation with higher standard errors. There is a cluster of low standard errors at low differences.

work [8] which calculated pulls from the difference of decay widths at a particular value of q^2.

The top four pull means for the SM can be found in Table 12. The top coefficient listed is $\mathrm{Re}(A_\perp^R)$ β with a pull mean of +10.10194. Although this coefficient has a low difference between the fit mean and signal, the standard error is low. It is therefore thought that the large pull mean is explained by the standard error rather than large differences between the fit values and the signal. This is also true for $\mathrm{Re}(A_\parallel^R)$ β. The other two coefficients, $\mathrm{Re}(A_0^L)$ α and $\mathrm{Re}(A_\parallel^R)$ α are listed in the table due to the large differences between their fit values and the signal. The difference in $\mathrm{Re}(A_\parallel^R)$ α is thought to be explained by the discrete symmetry as the fit mean is actually close to the absolute signal value, but with the opposite sign. All the coefficients listed in this table are real and there are no γ parameters. Two of the largest correspond to the $\mathrm{Re}(A_\parallel^R)$ amplitude.

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_\perp^R)$ β	+0.02730	+0.14706	+0.11976	0.01185	+10.10194
$\mathrm{Re}(A_0^L)$ α	+7.20276	+7.72792	+0.52516	0.06190	+8.48357
$\mathrm{Re}(A_\parallel^R)$ β	-0.00432	-0.10475	-0.10043	0.01205	-8.33771
$\mathrm{Re}(A_\parallel^R)$ α	-0.23538	+0.33127	+0.56665	0.07850	+7.21806

Table 12: Four largest pull means in the SM ensemble

The fit and pull distributions for the three largest values are shown in Fig. 14. It can be seen from this that there is no singular reason for why these coefficients are biased the way they are. $\mathrm{Re}(A_\perp^R)$ β shows a skewed distribution. The modal pull value is close to 0, but

the right tail causes the increase in the mean. This could mean that the bias is something to do with the fitting process. Re(A_0^L) α shows a symmetrical distribution so the bias seems more intrinsic to the coefficient itself but the reason for this is unknown. Re($A_\parallel^R$) β is also skewed but this time the modal value seems to overshoot 0 and correspond to a pull in a different direction. There is also a flattening to the pull distribution curve top.

The four lowest pull means for the SM ensemble are shown in Table 13. Apart from Re($A_\perp^R$) γ, the other three listed are imaginary. All of the coefficients listed correspond to ones with very low differences between the fit mean and the signal. Two of these correspond to the Im($A_\parallel^L$) amplitude.

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
Re($A_\perp^R$) γ	-7.14745	-7.14570	+0.00175	0.10741	+0.01632
Im($A_\parallel^L$) β	-0.00182	-0.00149	+0.00033	0.00686	+0.04858
Im($A_\parallel^L$) α	+0.00859	+0.00086	-0.00772	0.05081	-0.15195
Im(A_{00}^R) α	+1.00000	+0.99124	-0.00876	0.02580	-0.33943

Table 13: Four smallest pull means in the SM ensemble

The pull means for the NP ensemble are higher than the SM run as seen in Table 14. The highest is Re(A_0^L) α with a value of +18.56299. Even the lowest in this table, Re($A_\perp^L$) γ, with a value of -15.36866 still has a larger magnitude that the highest in the SM results. Like the SM ensemble all of the coefficients listed in this table correspond to real components, however unlike the SM run there are no β parameters listed. It is notable that Re(A_0^L) appears twice in the largest of the the NP ensemble. It is thought the Re(A_{00}^L) α appears due to its small standard error, whereas the other three have large differences between the mean fit and the signal. Everything in this table is a left amplitude.

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
Re(A_0^L) α	+5.88288	+6.97881	+1.09592	0.05904	+18.56299
Re(A_{00}^L) α	+1.00000	+1.19134	+0.19134	0.01095	+17.48009
Re(A_0^L) γ	+8.10140	+9.47912	+1.37772	0.08230	+16.73932
Re($A_\perp^L$) γ	-8.84114	-10.23768	-1.39653	0.09087	-15.36866

Table 14: Four largest pull means in the NP ensemble

The fit and pull distributions for the 3 largest pull means in this ensemble are shown in Fig. 15. These show the top three in this ensemble are all biased due to distribution skew as opposed to a shift to a symmetrical distribution. This implies that all these large pull means could be due to the fitting process and thus might be something that is possible to improve through tuning in the future.

Finally, the smallest pull means for the NP ensemble are shown in Table 15. All of these correspond to imaginary components and it is interesting that all three parameters

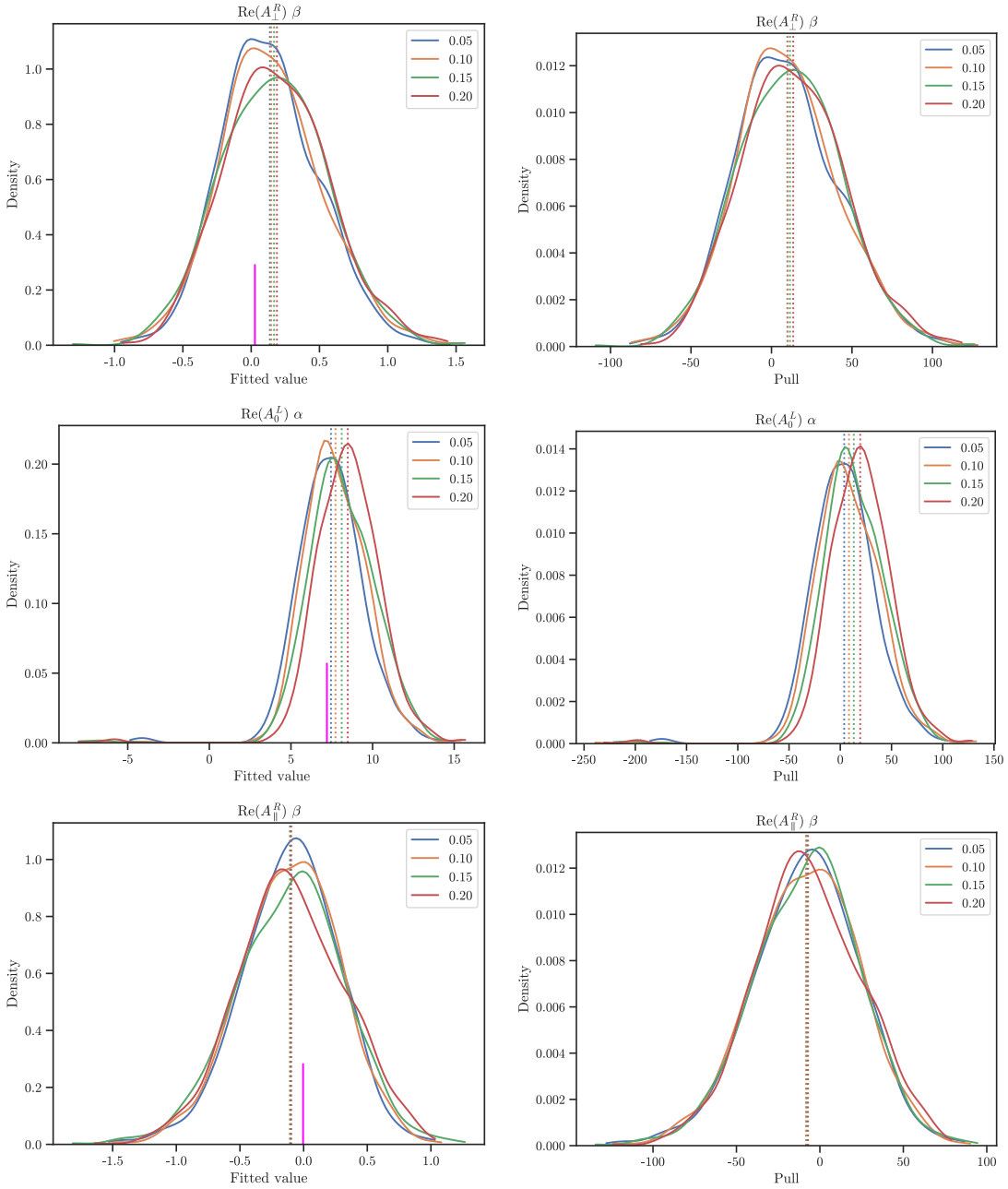

Figure 14: Plots of the three coefficients with the highest pulls using a 0.10 learning rate in the SM ensemble. Left plots show fit distributions and the right are pull distributions. Four different learning rates are shown but only 0.10 is discussed in the text. It can be seen from the shape of the Re($A_\perp^R$) β distributions (top) that the pull is to do with the skew of the distributions as the modal pull is actually near zero. Re(A_0^L) α (centre) doesn't have this skew and the pull seems more intrinsic to the coefficient. Re($A_\parallel^R$) β (bottom) shows an asymmetrical pull distribution for 0.10 and the modal pull value actually has the opposite sign.

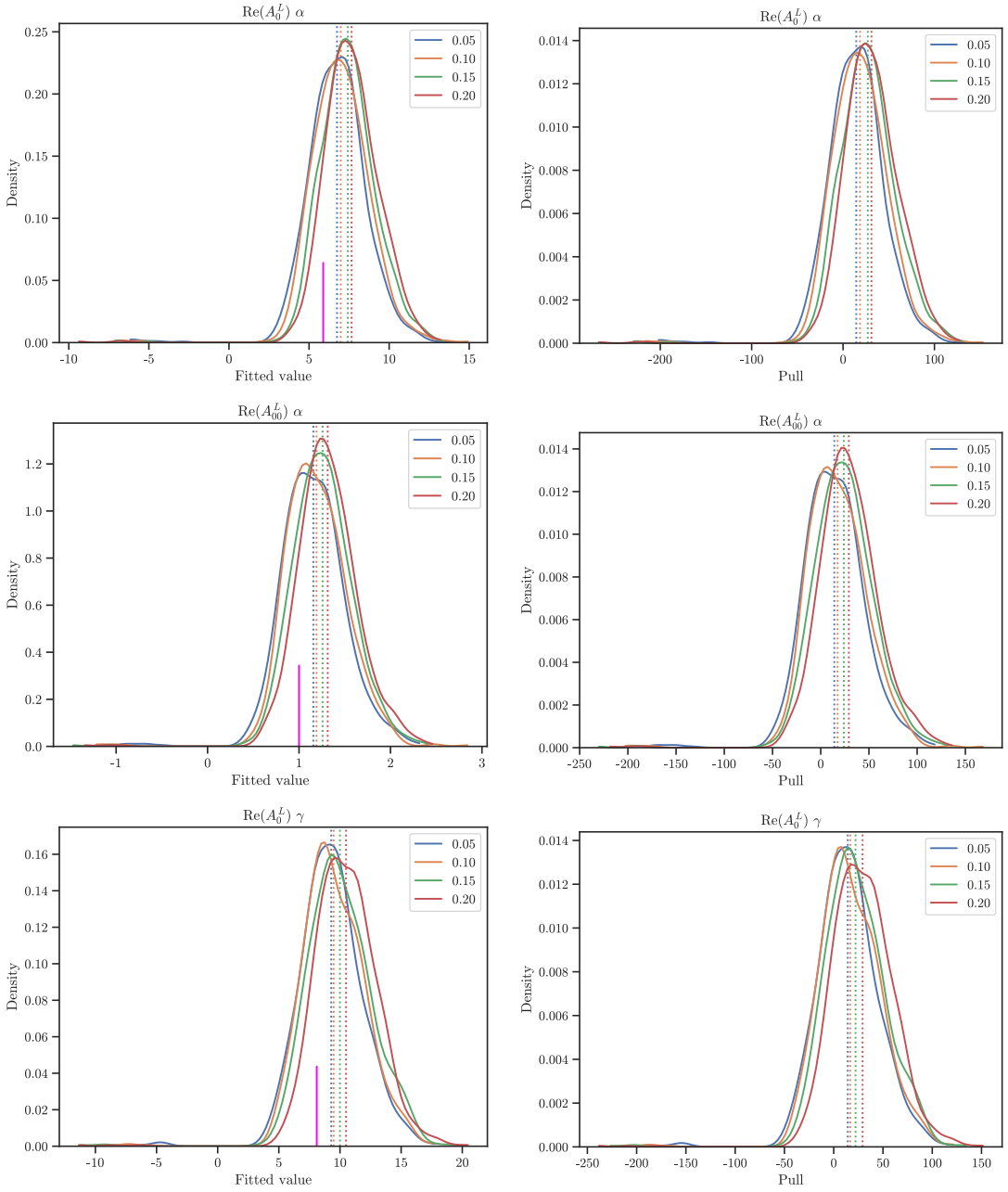

Figure 15: Plots of the three coefficients with the highest pulls using a 0.10 learning rate in the NP ensemble. Left plots show fit distributions and the right are pull distributions. Four different learning rates are shown but only 0.10 is discussed in the text. Re(A_0^L) α (top), Re(A_{00}^L) α (middle) and Re(A_0^L) γ (bottom) all have fairly smooth distributions and the pull mean seems related to the skew.

for the Im($A_\perp^L$) amplitude feature. All correspond to coefficients with small difference between the fit mean and signal.

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01742	-0.01651	+0.00091	0.01641	+0.05522
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00929	+0.01074	+0.00145	0.01041	+0.13898
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.11366	-0.13440	-0.02074	0.08069	-0.25702
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.04762	+0.00466	+0.05228	0.12424	+0.42078

Table 15: Four smallest pull means in the NP ensemble

An interesting pattern of the pull means emerges when plotting against the difference between the mean fit and signal values. This is shown in Fig. 16. The left plot is broken down by model and there appear to be two lines of correlation. One more vertical where coefficients with low differences have a range of pull means, and a more diagonal line where there is a loose correlation between the pull mean and the difference. The right plot is broken down by anzatz parameter and reveals more. Real β parameters clearly line up on the near vertical line. Real γ parameters show a strong link with the diagonal line. Real α shows points that line up on the diagonal line both left and right and the origin, but also on the vertical line above the origin. Imaginary α is fairly central but with some values in the upper-half of the vertical line. The other two imaginary coefficients are mostly constrained to the origin, having low differences and low pull means. No clear pattern was found by breaking this plot down by amplitude name.

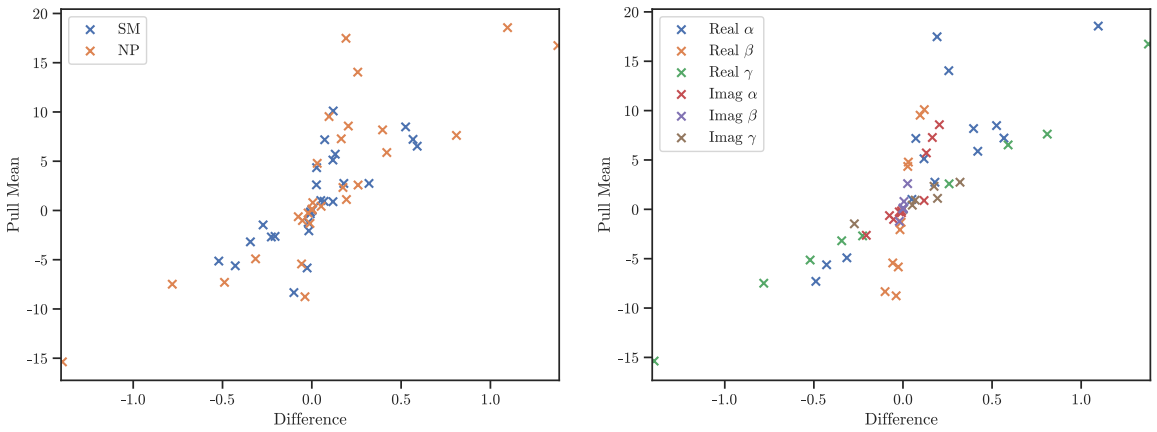

Figure 16: Plot of the pull mean versus the difference between the fit mean and signal for both the SM and NP ensembles. The left is broken down by model and the right by anzatz parameter. Two lines emerge. The first is almost vertical and corresponds to coefficients with low differences yet a range of pull means. The other line is diagonal and these coefficients show a rough correlation between the pull mean and difference.

3.4 Initialisation without randomisation

Another ensemble of 1000 fits in the SM but using the CURRENT_SIGNAL initialisation was generated and can be found in Appendix A.3. This run was done to see the if the bias introduced by the optimisation process could be reduced to indicate what the "truer" pulls would be. To stop the coefficients quickly moving away from their optimum values, the learning rate was reduced to 0.005. Other hyperparameters were left at their default values. As this initialisation scheme did not randomise coefficients, any fit that did not converge had to have the signal regenerated for the retry. This was because the restarted fit would have led to the same non-converging result. This regeneration is in contrast to the previous ensembles and happened for 12 fits out of the 1000.

The first interesting thing to point out is that this ensemble generally produced larger fit mean differences from the signal than the ensemble with randomisation. The highest difference in the last ensemble was Re(A_0^L) γ with a difference of +0.59006, and this coefficient is now the fourth largest with a difference of -0.88738. The largest difference in this ensemble was Re($A_\parallel^L$) γ with a difference of -1.31366. This was previously in 8th place with a difference of -0.27276. Both of theses plots can be seen in Fig. 17. The figure shows that although the fit means are further from the signal, the width of the distributions have narrowed. This narrowing was seen in the distribution for every coefficient.

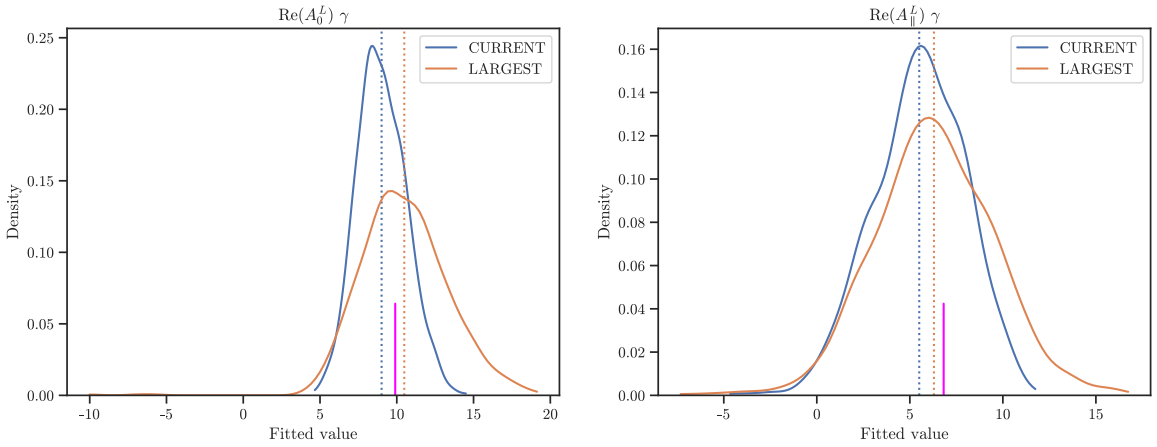

Figure 17: Comparison of the fit distributions for the TWICE_LARGEST_SIGNAL_SAME_SIGN (LARGEST) algorithm that has randomisation, and the CURRENT_SIGNAL (CURRENT) algorithm that does not. The magenta line shows the signal value. Re(A_0^L) γ (left) has the largest fit mean versus signal difference in the ensemble with randomisation and Re($A_\parallel^L$) γ (right) has the largest difference in the ensemble without. In both cases the scheme without randomisation produces fit means further from the signal, but the distribution is narrower.

The increase in difference is believed to be because the learning rate is still too high for this test and the coefficients quickly diverge from the optimum. Lowering it much further would however cause more signals to be rejected, in which case it wouldn't be fair to compare an ensemble of random signals to an ensemble with many selective ones. The

narrowing suggests that the standard error is something that could be improved by using a more intelligent initialisation algorithm.

Another interesting point is although the standard errors have been reduced, the 6 coefficients with the largest standard errors are the same and in the same order in both ensembles. These 6 coefficients include ones with large and small signals. This suggests that some coefficients inherently have larger standard errors, presumably because of the log likelihoods sensitivity to their change. It should be possible to explore this more in future by comparison of the derivatives of the log likelihood.

The pulls seen in this ensemble were higher than the previous run, but this is an expected consequence of having larger differences and smaller standard errors.

4 Discussion

4.1 Uncertainties

4.1.1 Random

The standard error is only a good estimate of the uncertainty for fit distributions that were symmetrical and Gaussian. For the 0.10 learning rate ensembles, these coefficients were:

- SM: $\mathrm{Im}(A_{\parallel}^{L})$ β, $\mathrm{Im}(A_{\parallel}^{R})$ α/γ, $\mathrm{Re}(A_{\perp}^{L})$ β/γ, $\mathrm{Im}(A_{\perp}^{L})$ β/γ and $\mathrm{Re}(A_{0}^{L})$ β.
- NP: $\mathrm{Im}(A_{\parallel}^{L})$ β/γ, $\mathrm{Re}(A_{\parallel}^{R})$ α/β, $\mathrm{Im}(A_{\parallel}^{R})$ $\alpha/\beta/\gamma$, $\mathrm{Re}(A_{\perp}^{L})$ β/γ, $\mathrm{Im}(A_{\perp}^{L})$ α/β, $\mathrm{Re}(A_{0}^{L})$ α, $\mathrm{Re}(A_{00}^{R})$ α and $\mathrm{Im}(A_{00}^{R})$ α.

The other coefficients either had skew in their distribution or had a non-Gaussian shape. The standard error is a bad estimate of the uncertainty for these cases.

It was attempted to understand and quantify the errors further by producing likelihood profiles by generating a single set of signal events and scanning one coefficient whilst letting the rest fit. An alternative method for finding the errors was also tried from inverting the Hessian matrix to find the covariance matrix. Both of these ideas failed to produce reasonable results because the likelihood curve produced near the optimum did not have a smooth shape. A alternative scan was performed by scanning each coefficient whilst fixing the rest at their signal values. This did produce a smooth likelihood. The reason for the non-smooth curve in the profile case is therefore believed to be due to needing to tune the learning rate specifically for these tests, but it wasn't possible to complete this in the time available. Performing these tests with a radically different learning rate would also fail to produce results applicable to the previous ensembles.

In order to quantify the uncertainties from the fitting process for each amplitude, confidence plots were generated. These were made by inputting a fitted α, β and γ for a fit result into the anzatz for an amplitude. Each amplitude anzatz for in an ensemble was then plotted over the q^2 range. The 68% and 95% confidence bands were calculated

from where 68% and 95% of anzatz points covered at each particular value of q^2. These correspond to 1 and 2 σ confidence respectively.

A comparison of the confidence plots for the real components of $A_{\parallel}^{L}$, $A_{\perp}^{L}$ and $A_{\perp}^{R}$ between the 0.10 SM and NP ensembles can be found in Fig. 18. These display a general trend for these ensembles that the SM mean fits are a good approximation to the signal, however the NP ensemble shows more divergence. The confidence bars show no noticeable trend in their size difference between the ensembles. The tightest 95% confidence level for a real component is seen for $A_{\perp}^{L}$ in both ensembles with a value of $+0.59/-0.58$ at $2.2\,\mathrm{GeV}^2/c^4$ in the SM ensemble and ± 0.59 at $2.9\,\mathrm{GeV}^2/c^4$ in the NP one. The largest 95% level for a real component is seen in A_{0}^{L} for both ensembles at $1\,\mathrm{GeV}^2/c^4$ with values of $+7.9/-7.0$ for the SM ensemble and $+7.0/-6.3$ in the NP one.

All coefficients have the largest confidence bars at $1\,\mathrm{GeV}^2/c^4$ which shrink to a minimum at some intermediate value, followed by expanding again towards $8\,\mathrm{GeV}^2/c^4$. The reason for this is believed to be that at low q^2, the errors are exaggerated by the combination of the uncertainty in α, and a large range of values that the uncertainty in γ produces. These are due to the fact that when the curves are plotted, below $1\,\mathrm{GeV}^2/c^4$ the q^{-2} dependence of γ greatly amplifies the uncertainty. At larger values of q^2 the uncertainty in β dominates and becomes multiplied due to the q^2 dependence of that parameter.

The imaginary components of $A_{\parallel}^{L}$, $A_{\parallel}^{R}$ and $A_{\perp}^{L}$ are compared in Fig. 19. These show a trend that the imaginary components in both ensembles have a fit mean close to the signal. This was shown earlier where imaginary components tend to have low signals and low fit mean differences. A notable difference from the real components is how wide the confidence bars are at $1\,\mathrm{GeV}^2/c^4$ before quickly shrinking before $\sim 2\,\mathrm{GeV}^2/c^4$. This is not as exaggerated with the real components. It is explained by a large spread of results for the α and γ components. The tightest imaginary 95% confidence level is seen for $A_{\perp}^{L}$ in both ensembles at $5.2\,\mathrm{GeV}^2/c^4$ with the values $+0.53/-0.52$ for the SM and ± 0.49 for NP. The largest imaginary 95% confidence level is seen with $A_{\parallel}^{R}$ at $1\,\mathrm{GeV}^2/c^4$ in both models with $+4.6/-4.7$ for the SM and $+5.2/-4.7$ for NP.

As the A_{00} coefficients only included α components, the confidence plots for these are of constant size across the q^2 range as would be expected.

To see if these uncertainties obeyed Poisson statistics, another ensemble was performed with a SM model but this time with a signal event count of 38,400. The full list of results for this can be found in Appendix A.4. The signal count is 16 times the signal of the 2,400 ensemble performed previously so a 4 factor reduction in confidence level sizes might be expected. This reduction was broadly seen in the imaginary components of $A_{\parallel}^{L}$ and $A_{\perp}^{L}$. The imaginary component of $A_{\parallel}^{R}$ was reduced by about a factor of 3. $\mathrm{Re}(A_{\perp}^{L})$ showed this change at the tightest point of $2.1\,\mathrm{GeV}^2/c^4$ but there was little change otherwise. $\mathrm{Re}(A_{\parallel}^{L})$ shows a 3 factor reduction at the tightest point, about a 2 factor reduction in q^2 values lower than this, but little change at higher q^2. The real components of $A_{\parallel}^{R}$ and $A_{\perp}^{R}$ show the opposite trend with about a 2 factor reduction at the minimum and higher values of q^2, but little change at lower values. There was about a 3 factor reduction seen for $\mathrm{Im}(A_{00}^{L})$, $\mathrm{Re}(A_{00}^{R})$ and $\mathrm{Im}(A_{00}^{R})$. The final two components, $\mathrm{Re}(A_{0}^{L})$ and $\mathrm{Re}(A_{00}^{L})$, showed

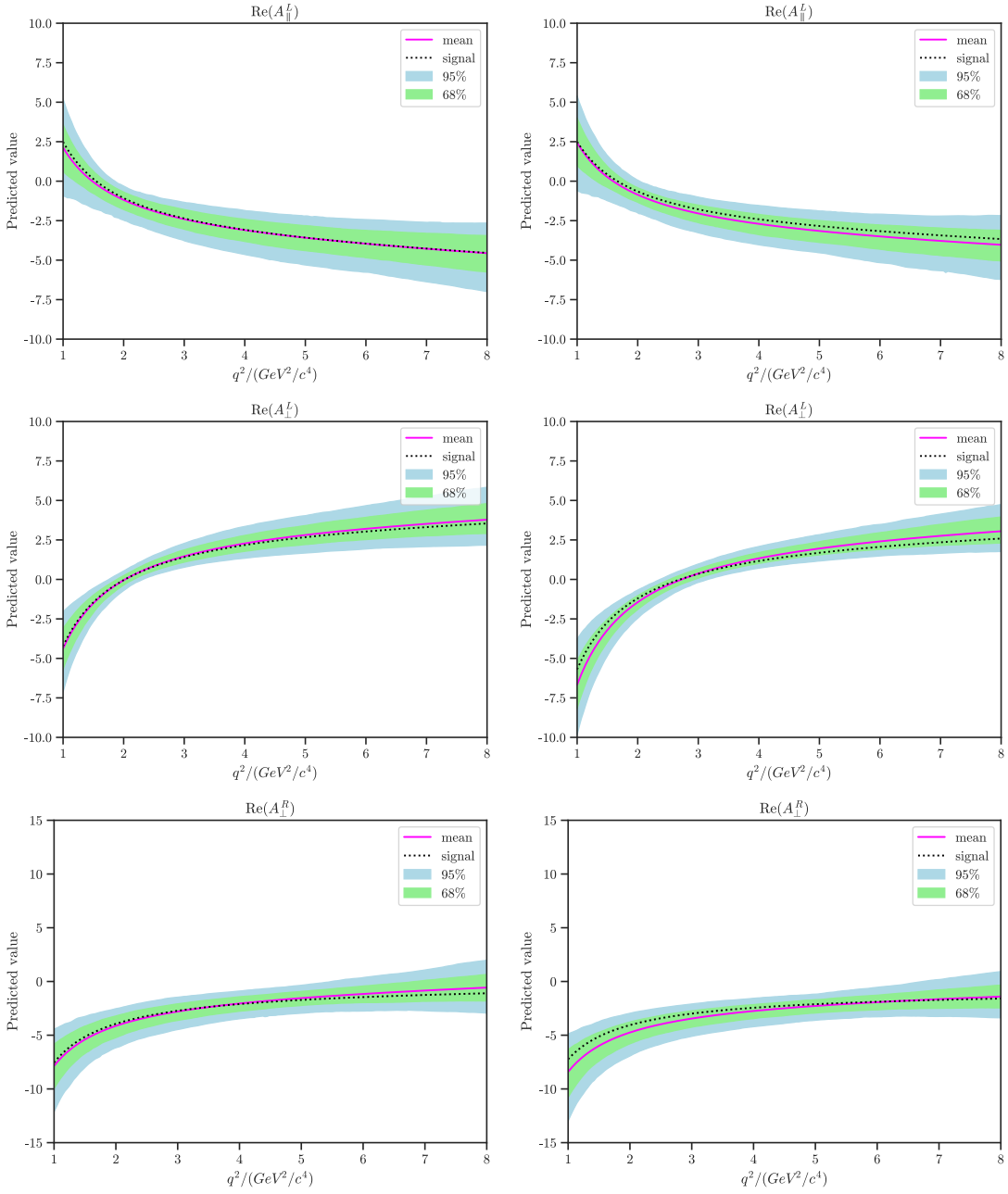

Figure 18: Comparison of confidence plots in the SM ensemble (left) and the NP ensemble (right). The real components of $A_\parallel^L$ (top), $A_\perp^L$ (middle) and $A_\perp^R$ (bottom) are shown. The SM plots show the fit mean tracking the signal better than the NP plots. Confidence bars are always narrower at intermediate values of q^2.

only a minimal effect from this signal increase.

Comparisons of the plots Im($A_\perp^L$) and Re($A_\parallel^R$) can be seen in Fig. 20. It was interesting that for some coefficients, increasing the signal changed the curve of the fit mean. Re($A_\parallel^R$)

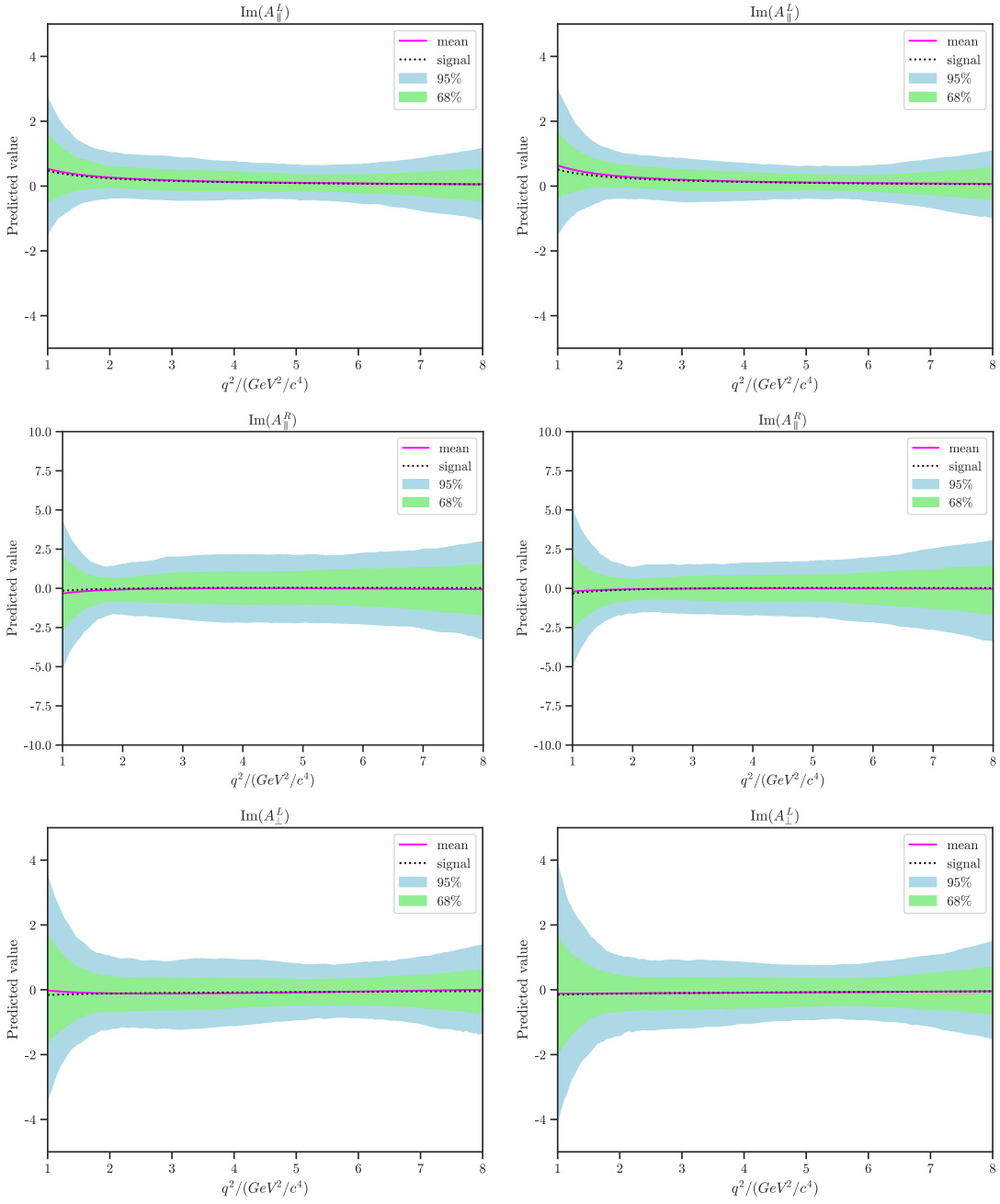

Figure 19: Comparison of confidence plots in the SM ensemble (left) and the NP ensemble (right). The imaginary components of $A_\parallel^L$ (top), $A_\parallel^R$ (middle) and $A_\perp^L$ (bottom) are shown. Both models show the fit mean tracking the signal equally well, and this is because the imaginary coefficients generally have low signal values and low fit mean differences from the signal. Unlike the real components, a large increase in uncertainty is seen at the q^2 lower limit of $1\,\mathrm{GeV}^2/c^4$.

diverged more from the signal at low q^2, $\mathrm{Re}(A_\parallel^L)$ and $\mathrm{Re}(A_\perp^R)$ were closer at higher q^2 but

further at lower q^2, and $\mathrm{Re}(A_\perp^L)$ and $\mathrm{Im}(A_{00}^R)$ were further away at all values of q^2.

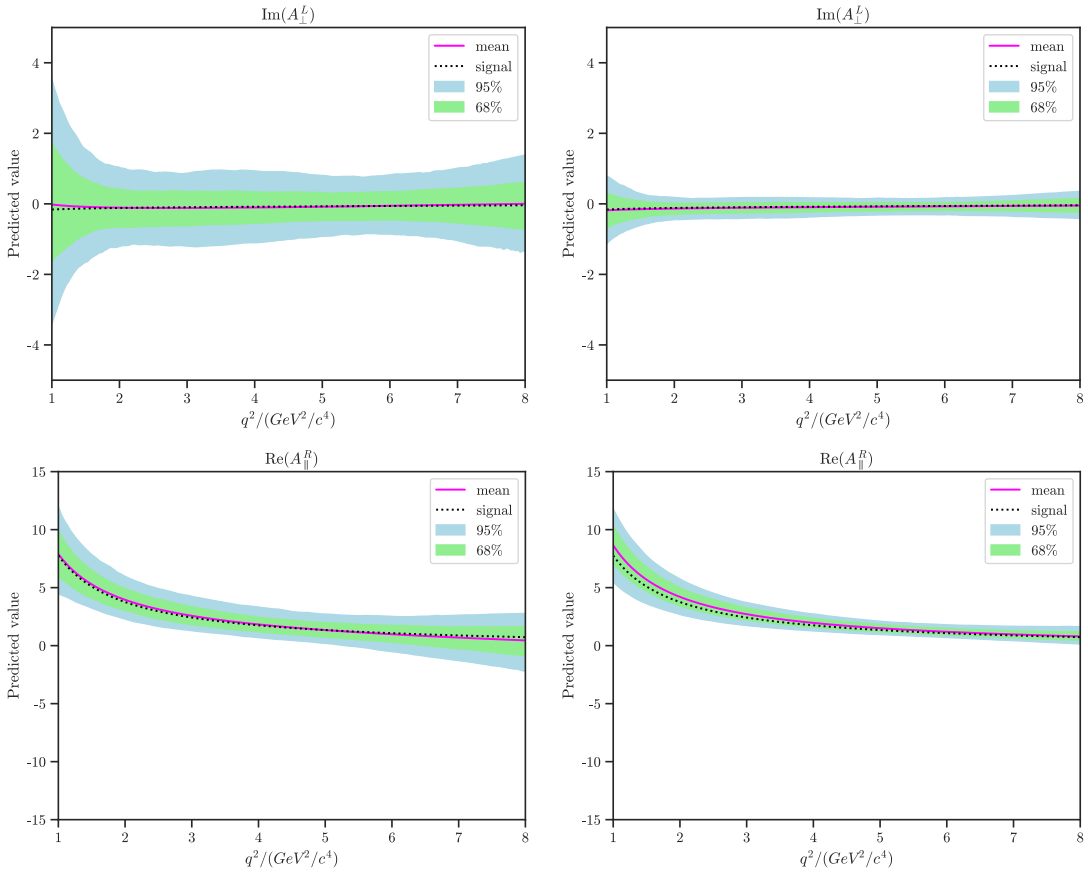

Figure 20: Comparison of confidence plots for a 2400 signal count (left) and a 38400 signal count (right). $\mathrm{Im}(A_\perp^L)$ (top) broadly shows a 4× reduction in the confidence level bands which would be expected from Poisson statistics. $\mathrm{Re}(A_\parallel^R)$ (bottom) shows a $\sim 2\times$ reduction at the tightest confidence point at at higher values of q^2, but little change at other values. The increase in signal also causes the $\mathrm{Re}(A_\parallel^R)$ fit mean to diverge from the signal curve.

A comparison was also done to see the effect of changing the learning rate on the confidence plots. Another SM ensemble was performed with a learning rate of 0.05 but with other machine learning hyperparameters left at their defaults. The full results of this can be found in Appendix A.6. The general trend was that reduced learning rate made either made no difference, made the fit mean slightly closer to the signal curve and/or provided a very small reduction in confidence bands. $\mathrm{Re}(A_\parallel^L)$ was interesting at it showed the reduced learning rate made the fit mean track the signal worse with a very small increase in confidence band size. Plots for $\mathrm{Re}(A_\parallel^L)$, and $\mathrm{Re}(A_0^L)$ which showed a better fit but no change to the confidence bands, can be found in Fig. 21.

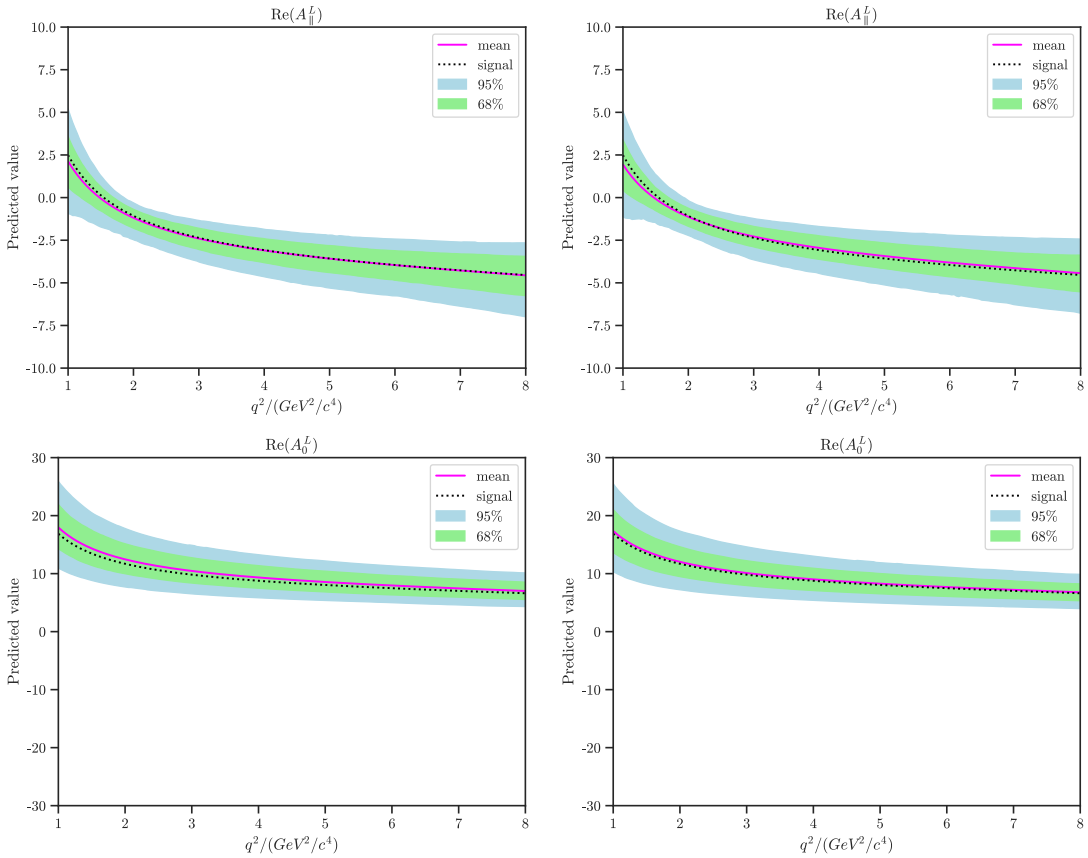

Figure 21: Comparison of confidence plots for a 0.10 learning rate (left) and a 0.05 learning rate (right). Re($A_\parallel^L$) (top) shows the reduced learning rate gives a slightly worse track of the fit mean to the signal and the confidence bands are slightly larger. This is unlike the other coefficients which either showed no change from the learning rate, or a small improvement with the reduced one. Re(A_0^L) (bottom) shows the reduced learning rate makes the fit mean track the signal better, but makes no noticeable difference to the confidence bands.

4.1.2 Systematic

There are systematic uncertainties introduced from the uncertainties on the mass and decay widths of the K^{*0} resonances that were used in the Breit-Wigner distributions. The $K^{*0}(892)$ is quoted in the 2018 PDG [12] as having a $895.55 \pm 0.20\,\text{MeV}$ mass and a $47.3 \pm 0.5\,\text{MeV}$ full width decay rate. The $K^{*0}(700)$ mass is given as $824 \pm 30\,\text{MeV}$ and decay width $478 \pm 50\,\text{MeV}$. To quantify the effect these had on the fit, another ensemble was performed with a SM signal. As the uncertainties were much smaller on the mass and decay width of $K^{*0}(892)$, as well as the masses of K, π and μ, these were neglected. The values for the $K^{*0}(700)$ mass and decay width were chosen as to provide the maximum change in the Breit-Wigner distributions previously calculated which was found to be by using the lower limit of the $K^{*0}(700)$ mass and the upper limit of the decay width. The new distribution values can be found in Table 16.

Integral	Value	
	Previous ensemble	This ensemble
$\int \lvert g_{K^{*0}(892)}\rvert^2 dm_{K\pi}$	0.888	0.888
$\int \lvert g_{K^{*0}(700)}\rvert^2 dm_{K\pi}$	0.296	0.248
$\int g_{K^{*0}(892)} g^*_{K^{*0}(700)} dm_{K\pi}$	0.640 + 0.00968 i	0.628 - 0.00295 i

Table 16: Values of the integrated relativistic Breit-Wigner distributions produced from using the 2018 PDG [12] central values for the $K^{*0}(700)$ mass and width (Previous ensemble) and by using the lower limit of mass and upper limit of decay width (This ensemble).

For this ensemble the learning rate was set as 0.10 with the other machine learning hyperparameters at their default values. The full list of results can be found in Appendix A.5. The difference in fit means, standard errors and the subsequently inferred systematic uncertainties are found in Table. 17. The systematic uncertainties associated with the $K^{*0}(700)$ are tentatively assumed to be $\pm$ the Δ Mean value. Reversing the direction of the $K^{*0}(700)$ mass and width limits would not be expected to produce uncertainties as high as the integrated BW values are closer to the originals for this case. Only 1000 fits were performed so these uncertainties may require more data to stabilise - this is something that should be followed up on in later work. The coefficients with the highest systematic errors are the α and γ components of $\mathrm{Re}(A_\parallel^R)$, $\mathrm{Re}(A_\perp^L)$, $\mathrm{Im}(A_\perp^L)$ and $\mathrm{Re}(A_\perp^R)$ as well as the α components of $\mathrm{Im}(A_\parallel^L)$ and $\mathrm{Im}(A_\parallel^R)$. The table shows there were no large differences in the standard errors between the two ensembles.

Systematic uncertainties have also been introduced from the process of producing the 3 parameter anzatz signal coefficients from the Wilson coefficients in each model. No attempt was made to quantify these uncertainties in this work.

4.2 Validity of results

The fact that the coefficients used in this study are known to produce incorrect differential decay rates, observable curves and S-wave fraction curve means the results in this paper are of mixed value. The fit means, pulls, and estimates of uncertainties are specific to the signal models used and therefore cannot be relied upon for physical answers. Later work will need to correct the coefficients in order to re-run the code to correct these results.

That being said, this paper has been able to prove the overall approach. Despite the models being non-physical, fitting code should be able to produce sensible results when working with arbitrary models. Failure to do so will end up with code that is biased towards a particular physics model. Lessons learnt about the importance of machine learning hyperparameters as well as the initialisation algorithm should be valuable to any later study. Patterns found in the fits for coefficients between the models should also prove true for other models. Furthermore the approximate width of the confidence bands is

Coefficient	Δ Mean	Δ Std. Err	Uncertainty
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-0.03884	+0.00138	$\pm$0.04
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	+0.00256	+0.00015	$\pm$0.003
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+0.06251	+0.00103	$\pm$0.06
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.02990	+0.00042	$\pm$0.03
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00546	+0.00004	$\pm$0.005
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	-0.10561	-0.00009	$\pm$0.1
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.12190	+0.00274	$\pm$0.1
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	+0.01874	+0.00059	$\pm$0.02
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+0.24886	+0.00059	$\pm$0.2
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	-0.06488	+0.00338	$\pm$0.1
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	+0.00751	+0.00012	$\pm$0.01
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	+0.23406	+0.00420	$\pm$0.2
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+0.10450	-0.00124	$\pm$0.1
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	-0.00994	-0.00025	$\pm$0.01
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-0.12557	-0.00208	$\pm$0.1
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	+0.19896	-0.00156	$\pm$0.2
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	-0.02848	-0.00014	$\pm$0.03
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.24653	-0.00168	$\pm$0.2
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.14743	-0.00082	$\pm$0.1
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.01907	-0.00050	$\pm$0.02
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	+0.19144	+0.00299	$\pm$0.2
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+0.08164	+0.00207	$\pm$0.1
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.00581	+0.00011	$\pm$0.01
$\mathrm{Re}(A_{0}^{L})\ \gamma$	-0.02808	-0.00152	$\pm$0.03
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+0.01270	+0.00029	$\pm$0.01
$\mathrm{Im}(A_{00}^{L})\ \alpha$	-0.02703	+0.00012	$\pm$0.03
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+0.02987	+0.00102	$\pm$0.03
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+0.09055	+0.00100	$\pm$0.1

Table 17: Differences between the fit means and standard errors, and the subsequent uncertainty between an ensemble that used extreme values of $K^{*0}(700)$ mass and decay width versus one that used the central values. The systematic uncertainty caused by the uncertainty in $K^{*0}(700)$ values is tentatively taken as $\pm$ the Δ Mean.

believed to be inherent to the approach in this paper and not dependant on the coefficients chosen.

4.3 Fit improvements

There is the potential to improve the fit quality and reduce the confidence bands. As previously discussed, various machine learning hyperparameters have an effect on the quality of the solution. This work only looked at each hyperparameter individually at various arbitrarily chosen values. A more rigorous treatment would involve searching the parameter space of all hyperparameters at the same time to find optimum values. Instead of doing it manually, this would be better performed as a meta-machine learning operation. The HParams [22] functionality for the Tensorboard dashboard may be of value here.

Another improvement that could be performed is optimising the fit coefficient initialisation values. As seen previously, the choice of initialisation can produce large differences in the fit mean and the width of the distribution. The TWICE_LARGEST_SIGNAL_SAME_SIGN algorithm was created with the intention that fits shouldn't rely on prior knowledge of the particular signal model used, but it causes a large range of starting values for coefficients with large signal values. Say for example a coefficient was 6 in the SM and 7 with the NP model. The range 0 to 14 would be used to start that coefficient. A smaller range of for example, $-\Delta 1$ the smallest coefficient and $+\Delta 1$ the largest coefficient may be more appropriate whilst fulfilling the stated goal of not relying on a particular model.

It may be also worth revisiting the choice of optimiser algorithm used. The AMSGrad algorithm uses momentum that has the effect of not always finding the optimum value. By implementing normalisation of the coefficients, it is possible that a simpler algorithm like SGD or RMSProp may allow quick convergence and produce less biased results.

For all of these ideas there are two important points to bear in mind. The first is that tuning should not be performed against one particular signal model. At minimum the ideal solution should perform equally well on the SM and NP models used in this study and it may be also worth introducing more models to ensure they perform well too. Additionally care must be taken to observe the shape of fit distributions. An ideal solution would produce not only a fit mean equalling the signal, but also a symmetrical fit distribution with no skew. Failure to consider this would result in coefficient biases from the method itself.

4.4 Performance improvements

The total time for each fit is largely influenced by the number of minimisation steps needed, and these are dependent on the machine learning hyperparameters and coefficient initialisation algorithm previously discussed. A smaller fit speedup could be achieved by removing the signal generation at the start of each fit which is a CPU intensive blocking operation. By pre-computing signal events, or by generating them in parallel to the

optimisation process so they are ready for the next fit, a couple of seconds would be shaved off each fit by keeping the GPU better utilised.

As previously stated when giving the definition of the angular observables, the terms β_μ and m_μ were left in. This is unnecessary when the massless limit is assumed ($q^2 \gg 4m_\mu^2$). The m_μ terms could be dropped and $\beta_\mu \to 1$ could be set [6]. Furthermore under those simplifications, two observables can be related to the others by the relations $J_{2c} = -J_{1c}$ and $J_{2s} = J_{1s}/3$ [8]. These simplifications and relationships would reduce the amount of computation needed per fit step, although only by a modest amount.

Improvements to the underlying code may be possible which could improve the time per minimisation step using the profiler in Tensorboard to study where the processor time is spent. This was used earlier in the project to provide an order of magnitude speed increase by studying operations that were running on the CPU and changing the code so they stayed on the GPU. Another potential code speedup would be to try to better optimise Tensorflow's autograph [23] functionality that converts Python code to Tensorflow graph code. This is currently implemented on the optimiser gradient calculation function only, but further improvements may be possible by moving this to alternate functions.

It isn't expected that any of these methods will result in a large improvement in performance, however they may still be worthwhile if the code used in this study is reused in future.

The ensembles in this study were run in series using a single GPU. The code will support choosing a running on a specific GPU so multiple ensembles could be run in parallel on a multi-GPU machine or multiple machines.

4.5 NP sensitivity

A statistical test was performed to test how sensitive this work was to NP. To do this a signal was repeatedly generated and two fits were performed to it. The first fit fixed the P-wave coefficients to the NP signal values (test hypothesis), and the second fit fixed them to the SM values (null hypothesis). The S-wave coefficients were left as trainable as placeholder values were used in this study. The previous study also left them as trainable, but because the values they used were less understood experimentally [8]. For each signal the Q test statistic was then calculated which was defined as

$$\mathrm{Q} = 2(\mathrm{NLL}_{test} - \mathrm{NLL}_{null}) \tag{27}$$

with NLL the negative log likelihood for the test or null hypothesis. This was repeated 5000 times for an SM signal, and 5000 times for a NP signal. To calculate the sensitivity the fraction of fits with $Q^{SM} \leq \bar{Q}^{NP}$ was counted where Q^{SM} was the Q statistic for the SM signal and $\bar{Q}^{NP}$ was the median for the NP signal. To calculate this properly would require a huge number of fits, so the Q^{SM} fraction distribution was extended by an approximation of a Gaussian fit to the SM results. This test neglects the systematic uncertainties previously discussed. The learning rate was set as 0.10 for this test, ϵ as 10^{-5} and the β hyperparameters were left at their default values. These hyperparameters

were chosen to try and improve the test performance as much as possible in the run time available.

This test was first performed for a signal event count of 600 which can be seen in Fig. 22. This number was what was chosen in the previous work [8] and was based on the Run-I number of signal events. That paper used an upper q^2 range of $6\,\mathrm{GeV}^2/c^4$ and the $8\,\mathrm{GeV}^2/c^4$ used in this paper would result in a higher signal count, however the number was kept the same for a fairer comparison of results.

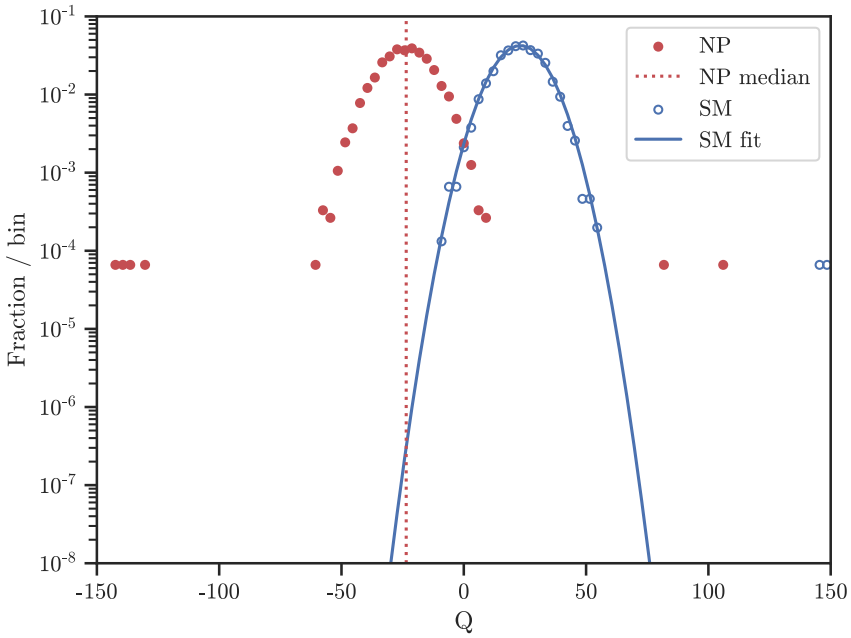

Figure 22: Distribution of 5,000 Q test statistics over 100 bins for an ensemble of fits to 600 signal events in the NP model (red) and the SM (blue). The Q statistic was defined as $2(\mathrm{NLL}_{\mathrm{test}} - \mathrm{NLL}_{\mathrm{null}})$ where the test and null labels correspond to fixing the P-wave coefficients to the NP and SM signal values respectively. The red dotted line is the median of NP statistics, and the blue line is a Gaussian fit to SM statistics. A confidence level of 4.9σ for being able to differentiate the two models was calculated from where the lines intersect. The bins were only used for plotting and not for calculations. The outliers have negligible importance due to the logarithmic scale.

It is interesting to note that on the plot, various outlier results are seen for both signal models. These correspond to individual or extremely small numbers however so should only negligibly affect the final result.

A 4.9 σ confidence level was found. This is unfavourable to the CP-averaged 6.5 σ found in the previous study [8], especially considering this work excluded the background. It should be pointed out though that the coefficient values used in this paper are known to be wrong, so a direct comparison of results isn't possible. No attempt was made to compare this result to a binned fit approach due to time constraints.

The test was then performed for a signal count of 2400 which matched the expected Run-II count used in the previous work [8] but again was not adjusted for the increase in q^2 range. This can be seen in Fig. 23. In this run a 10.0 σ confidence level was found for being able to differentiate NP from the SM. On this plot there are many more outliers with higher fractions, especially in the NP model. Given the logarithmic scale, these are still small fractions but will certainly have more effect than in the 600 signal count case.

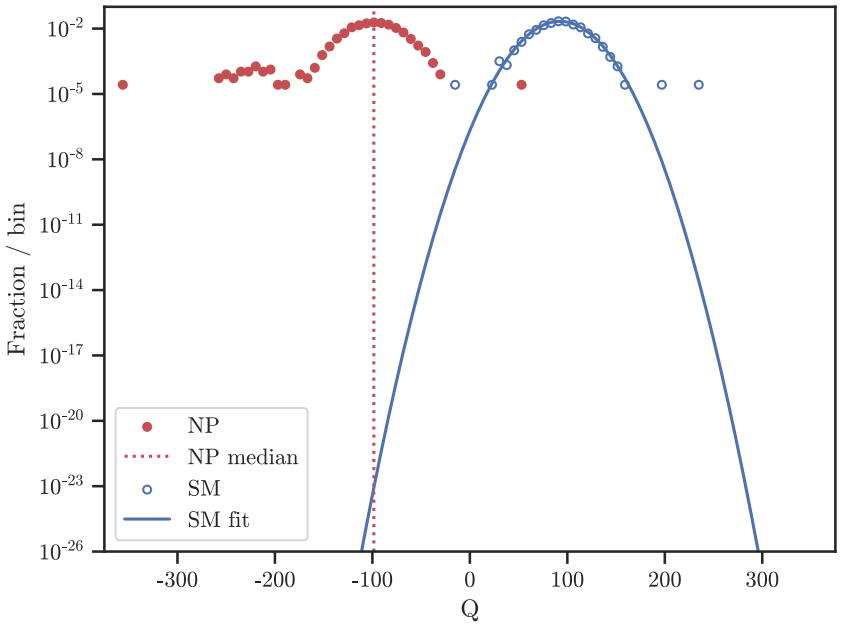

Figure 23: Distribution of 5,000 Q test statistics over 100 bins for an ensemble of fits to 600 signal events in the NP model (red) and the SM (blue). The red dotted line is the median of NP statistics, and the blue line is a Gaussian fit to SM statistics. A confidence level of 10.0σ for being able to differentiate the two models was calculated from where the lines intersect. Compared to the 600 event count plot in Fig. 22, the outliers are more significant.

4.6 Comparison with previous work

Two notable simplifications were made in this study compared to the previous work. Firstly the background was neglected. Secondly only the B^0 was considered and nothing was done to compare CP-averaged versus CP-asymmetric observables. This should have led to fits converged easier.

The pulls in this study were calculated in a different way so cannot be compared directly, however it is possible to compare the confidence plots for coefficients directly that implicitly include this. Fig. 24 contains a comparison for the $\mathrm{Re}(A_\perp^L)$ and $\mathrm{Re}(A_\perp^R)$ amplitudes in SM ensembles. What can be seen is that despite the values for the coefficients in this study are known to be incorrect, the overall shape of the plots are similar. The fit means show a

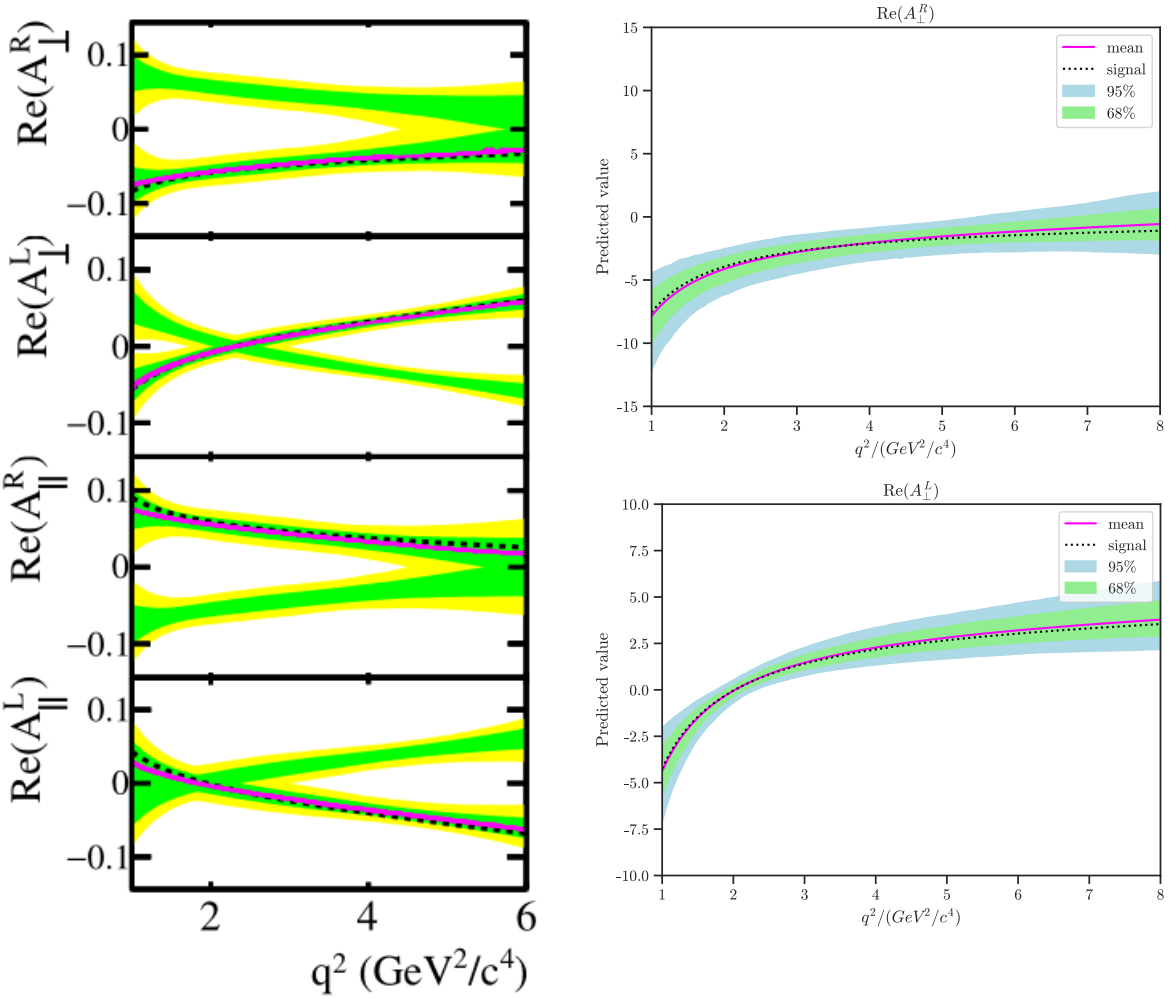

Figure 24: Comparison of SM confidence plots produced by the previous work [8] (left) and $\mathrm{Re}(A_\perp^R)$ (top right) and $\mathrm{Re}(A_\perp^L)$ (bottom right) produced by this study (0.10 learning rate). The magenta line represents the fit mean, green is the 68% confidence band, yellow/blue the 95% and the dotted line shows the signal values. The left plots show two bands due to the discrete symmetries that were mostly removed in this study by choice of the initialisation algorithm. Apart from the scale differences, $\mathrm{Re}(A_\perp^R)$ shows a similar result and level of relative confidence. $\mathrm{Re}(A_\perp^L)$ has a similar fit mean and confidence at lower q^2, but the confidence bands in this work are noticeably larger after $\sim 2\,\mathrm{GeV}^2/c^4$.

similar level of divergence from the signal over the $1\,\mathrm{GeV}^2/c^4 < q^2 < 6\,\mathrm{GeV}^2/c^4$ range used in the previous study. The previous work's plots include the discrete symmetries that were mostly removed from the study by choice of the initialisation algorithm. For that reason, the single lines from this work would be expected to display wider confidence bands than would be seen in a two line plot. The confidence bands for $\mathrm{Re}(A_\perp^R)$ are fairly similar in size with this study performing slightly better near $1\,\mathrm{GeV}^2/c^4$ and the previous study slightly better at higher values. $\mathrm{Re}(A_\perp^L)$ shows a similar confidence level below $2\,\mathrm{GeV}^2/c^4$ but this

work has larger confidence band at higher energies. Similar patterns are seen in most the rest of the confidence plots with the shape being similar, the mean tracking the signal to a similar level, and the confidence bands looking similar or being slightly larger, especially at higher values of q^2. The difference at higher energies would suggest this study was unable to constrain the β parameters as well. A notable exception was $\mathrm{Re}(A_0^L)$ which displays a different shape, but this is not surprising considering the coefficients used in this study are known to have errors.

Uncertainties in the previous study were derived from producing a likelihood profile. As previously stated, that wasn't possible to do in this work and so no comparison can be made of errors.

5 Conclusion

This study has done what it set out to do in proving that Google Tensorflow could be a useful tool in fitting to angular observables for $B^0 \to K^{*0}\mu^+\mu^-$ decays. The amplitude fit mean provides a good fit to the signal for only a modest signal size. The uncertainty bars produced by the amplitudes weren't insignificant, but similar sizes were seen in the previous study. It should be possible to improve on what was done in this paper and reduce them in later studies by further tuning. This work suffered from large pull means for some coefficients. It may be possible to address this by either further tuning, or by making another optimiser algorithm work with a technique like coefficient normalisation.

Deviations from lepton-universality in $B \to K\ell\ell$ decays have been found, as well as SM tension in the angular observables of $B^0 \to K^{*0}\mu^+\mu^-$. Constraining the Wilson coefficients is necessary to motivate the hunt for new physics that could explain these results, whether it be in the form of a U_1 leptoquark, new scalar bosons, or something else. To take this work toward that goal, it will be necessary to run this work on real world data and perform a scan of the Wilson coefficients. The best-fit and error matrix can them be used to provide constraints as discussed in the Bayesian analysis of [24].

Before this work can be run on real data however, there are certain points that need addressing. Firstly the coefficients used in this study will need correcting to physical values. It will be possible to test these are correct by comparing the differential decay rate, S-wave fraction, and observable plots against what is expected. Secondly this work will need expanding to also include $\overline{B}^0$ decays separately. This will require the inclusion of CP-averaged and CP-asymmetric observables. Finally the background will need adding. If the fits are still able to converge then Google Tensorflow may provide a useful tool to constrain NP searches in the future.

Acknowledgements

Credit to Ulrik Egede, Mitesh Patel and Konstantinos Petrisdis for their previous work [8] that was built on in this paper. Many thanks to Mitesh Patel and Mark Smith of Imperial

College for the numerous helpful chats. Additional thanks to Mark Smith for generating the signal coefficients from flavio that were used in this study.

Declaration of Work

All code that was written for this project and all analysis was completed solely by myself. No pre-existing code for this method was used.

Appendix

A Results

A.1 SM

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-4.17810	-3.99891	+0.17919	0.06524	+2.74672
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	-0.15184	-0.16856	-0.01672	0.00809	-2.06610
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+6.81832	+6.29739	-0.52094	0.10144	-5.13558
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.00859	+0.00086	-0.00772	0.05081	-0.15195
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00182	-0.00149	+0.00033	0.00686	+0.04858
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	+0.46607	+0.53478	+0.06871	0.07358	+0.93378
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.23538	+0.33127	+0.56665	0.07850	+7.21806
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	-0.00432	-0.10475	-0.10043	0.01205	-8.33771
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+8.00375	+7.65988	-0.34386	0.10793	-3.18590
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	+0.16564	+0.28405	+0.11841	0.13308	+0.88970
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01310	-0.03372	-0.02062	0.01694	-1.21727
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	-0.30668	-0.57944	-0.27276	0.18585	-1.46762
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+3.88641	+3.93645	+0.05004	0.04978	+1.00536
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	+0.08527	+0.11211	+0.02685	0.00618	+4.34479
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-8.19745	-8.42401	-0.22657	0.08420	-2.69079
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.09505	-0.30081	-0.20576	0.07814	-2.63319
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00793	+0.03393	+0.02599	0.00998	+2.60340
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.07297	+0.24708	+0.32005	0.11652	+2.74662
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.42358	-0.85217	-0.42859	0.07631	-5.61654
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.02730	+0.14706	+0.11976	0.01185	+10.10194
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	-7.14745	-7.14570	+0.00175	0.10741	+0.01632
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+7.20276	+7.72792	+0.52516	0.06190	+8.48357
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.22782	-0.25409	-0.02628	0.00450	-5.83660
$\mathrm{Re}(A_{0}^{L})\ \gamma$	+9.89863	+10.48869	+0.59006	0.09043	+6.52481
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+1.00000	+1.07207	+0.07207	0.01003	+7.18181
$\mathrm{Im}(A_{00}^{L})\ \alpha$	+1.00000	+1.13120	+0.13120	0.02296	+5.71306
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+1.00000	+1.11802	+0.11802	0.02304	+5.12160
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+1.00000	+0.99124	-0.00876	0.02580	-0.33943

Table 18: Values of the signal, fit mean, difference between the fit mean and signal, standard error and pull mean for an ensemble of 1000 fits using SM signal. Only trained coefficients are included. A learning rate of 0.10 was used for this data with β_1, β_2 and ϵ left at their default values.

A.2 NP

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-3.42775	-3.74269	-0.31494	0.06413	-4.91102
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	-0.12410	-0.13436	-0.01026	0.00782	-1.31241
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+6.04528	+6.30423	+0.25895	0.10023	+2.58346
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.00934	-0.04313	-0.05247	0.05130	-1.02293
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00199	+0.00349	+0.00548	0.00689	+0.79509
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	+0.50341	+0.67775	+0.17434	0.07475	+2.33228
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.25087	+0.16884	+0.41971	0.07136	+5.88169
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	-0.00518	-0.06252	-0.05733	0.01056	-5.42929
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+8.63674	+9.44652	+0.80978	0.10630	+7.61804
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	+0.22209	+0.14636	-0.07573	0.11937	-0.63439
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01742	-0.01651	+0.00091	0.01641	+0.05522
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	-0.52807	-0.33425	+0.19382	0.17437	+1.11149
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+3.06464	+3.46083	+0.39619	0.04843	+8.18077
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	+0.07852	+0.10878	+0.03026	0.00632	+4.78584
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-8.84114	-10.23768	-1.39653	0.09087	-15.36866
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.11366	-0.13440	-0.02074	0.08069	-0.25702
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00929	+0.01074	+0.00145	0.01041	+0.13898
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.04762	+0.00466	+0.05228	0.12424	+0.42078
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.93327	-1.42217	-0.48891	0.06699	-7.29802
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.01687	+0.11321	+0.09635	0.01011	+9.53352
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	-6.31856	-7.09980	-0.78124	0.10435	-7.48699
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+5.88288	+6.97881	+1.09592	0.05904	+18.56299
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.18442	-0.22278	-0.03835	0.00438	-8.75991
$\mathrm{Re}(A_{0}^{L})\ \gamma$	+8.10140	+9.47912	+1.37772	0.08230	+16.73932
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+1.00000	+1.19134	+0.19134	0.01095	+17.48009
$\mathrm{Im}(A_{00}^{L})\ \alpha$	+1.00000	+1.20432	+0.20432	0.02383	+8.57276
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+1.00000	+1.25710	+0.25710	0.01831	+14.04339
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+1.00000	+1.16480	+0.16480	0.02265	+7.27481

Table 19: Values of the signal, fit mean, difference between the fit mean and signal, standard error and pull mean for an ensemble of 1000 fits using NP signal. Only trained coefficients are included. A learning rate of 0.10 was used for this data with β_1, β_2 and ϵ left at their default values.

A.3 SM (CURRENT_SIGNAL initialisation)

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-4.17810	-3.48769	+0.69042	0.04872	+14.17158
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	-0.15184	-0.14320	+0.00864	0.00665	+1.29985
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+6.81832	+5.50466	-1.31366	0.07747	-16.95780
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.00859	-0.00271	-0.01129	0.04085	-0.27641
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00182	+0.00105	+0.00287	0.00553	+0.51847
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	+0.46607	+0.42230	-0.04377	0.05996	-0.72999
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.23538	+0.21925	+0.45463	0.06592	+6.89640
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	-0.00432	-0.08003	-0.07571	0.01024	-7.39654
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+8.00375	+6.74462	-1.25912	0.08682	-14.50321
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	+0.16564	+0.10096	-0.06468	0.10320	-0.62682
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01310	-0.01235	+0.00075	0.01312	+0.05714
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	-0.30668	-0.23694	+0.06974	0.14376	+0.48513
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+3.88641	+3.46804	-0.41836	0.03574	-11.70475
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	+0.08527	+0.08933	+0.00406	0.00484	+0.84022
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-8.19745	-7.37584	+0.82161	0.05486	+14.97670
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.09505	-0.21387	-0.11882	0.06192	-1.91907
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00793	+0.02433	+0.01639	0.00802	+2.04531
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.07297	+0.09920	+0.17217	0.09286	+1.85405
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.42358	-0.70096	-0.27738	0.06235	-4.44865
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.02730	+0.11878	+0.09148	0.00930	+9.83694
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	-7.14745	-6.18761	+0.95984	0.08116	+11.82621
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+7.20276	+6.72000	-0.48276	0.03623	-13.32365
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.22782	-0.22118	+0.00663	0.00395	+1.67838
$\mathrm{Re}(A_{0}^{L})\ \gamma$	+9.89863	+9.01125	-0.88738	0.04994	-17.76833
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+1.00000	+0.92995	-0.07005	0.00649	-10.79276
$\mathrm{Im}(A_{00}^{L})\ \alpha$	+1.00000	+0.95549	-0.04451	0.01731	-2.57165
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+1.00000	+0.96886	-0.03114	0.01904	-1.63536
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+1.00000	+0.91536	-0.08464	0.02044	-4.14169

Table 20: Values of the signal, fit mean, difference between the fit mean and signal, standard error and pull mean for an ensemble of 1000 fits using SM signal and the CURRENT_SIGNAL coefficient initialisation algorithm. Only trained coefficients are included. A learning rate of 0.005 was used for this data with β_1, β_2 and ϵ left at their default values.

A.4 SM (8× signal events)

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-4.17810	-4.64026	-0.46215	0.03668	-12.59839
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	-0.15184	-0.16386	-0.01202	0.00224	-5.36777
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+6.81832	+7.54623	+0.72790	0.05961	+12.21195
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.00859	+0.03246	+0.02387	0.01368	+1.74460
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00182	-0.00435	-0.00253	0.00185	-1.36203
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	+0.46607	+0.48378	+0.01770	0.01948	+0.90879
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.23538	-0.20272	+0.03266	0.02335	+1.39878
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	-0.00432	-0.01364	-0.00933	0.00361	-2.58119
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+8.00375	+8.86316	+0.85942	0.06047	+14.21236
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	+0.16564	+0.22811	+0.06246	0.04140	+1.50876
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01310	-0.02142	-0.00832	0.00608	-1.36838
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	-0.30668	-0.37708	-0.07040	0.05378	-1.30917
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+3.88641	+4.29152	+0.40511	0.03228	+12.54946
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	+0.08527	+0.09692	+0.01165	0.00170	+6.86515
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-8.19745	-9.05947	-0.86202	0.06612	-13.03756
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.09505	-0.10800	-0.01295	0.02134	-0.60672
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00793	+0.00874	+0.00080	0.00273	+0.29432
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.07297	-0.08403	-0.01106	0.03193	-0.34646
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.42358	-0.50394	-0.08036	0.02210	-3.63572
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.02730	+0.03988	+0.01258	0.00332	+3.78932
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	-7.14745	-7.95182	-0.80437	0.05611	-14.33556
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+7.20276	+7.98839	+0.78563	0.05795	+13.55792
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.22782	-0.25183	-0.02402	0.00213	-11.25005
$\mathrm{Re}(A_{0}^{L})\ \gamma$	+9.89863	+11.00261	+1.10398	0.08017	+13.77054
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+1.00000	+1.11284	+0.11284	0.00824	+13.69949
$\mathrm{Im}(A_{00}^{L})\ \alpha$	+1.00000	+1.10615	+0.10615	0.00972	+10.92271
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+1.00000	+1.12837	+0.12837	0.00824	+15.58777
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+1.00000	+1.12222	+0.12222	0.00923	+13.24478

Table 21: Values of the signal, fit mean, difference between the fit mean and signal, standard error and pull mean for an ensemble of 1000 fits using SM signal. Only trained coefficients are included. The signal count was increased to 38400 which is 8× the 2400 used in the other ensembles. A learning rate of 0.10 was used for this data with β_1, β_2 and ϵ left at their default values.

A.5 SM (Extreme $K^{*0}(700)$ mass and width)

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-4.17810	-4.03775	+0.14035	0.06662	+2.10669
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	-0.15184	-0.16600	-0.01416	0.00824	-1.71762
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+6.81832	+6.35990	-0.45842	0.10247	-4.47389
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.00859	+0.03076	+0.02218	0.05123	+0.43291
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00182	-0.00695	-0.00512	0.00690	-0.74249
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	+0.46607	+0.42917	-0.03690	0.07349	-0.50213
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.23538	+0.20937	+0.44475	0.08124	+5.47457
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	-0.00432	-0.08601	-0.08169	0.01264	-6.46198
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+8.00375	+7.90874	-0.09500	0.11525	-0.82429
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	+0.16564	+0.21917	+0.05353	0.13646	+0.39230
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01310	-0.02621	-0.01311	0.01706	-0.76833
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	-0.30668	-0.34538	-0.03870	0.19005	-0.20362
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+3.88641	+4.04095	+0.15455	0.04854	+3.18412
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	+0.08527	+0.10217	+0.01690	0.00593	+2.85187
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-8.19745	-8.54958	-0.35214	0.08212	-4.28800
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.09505	-0.10185	-0.00680	0.07658	-0.08873
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00793	+0.00545	-0.00249	0.00984	-0.25276
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.07297	+0.00055	+0.07352	0.11484	+0.64016
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.42358	-0.99960	-0.57601	0.07549	-7.63026
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.02730	+0.16613	+0.13883	0.01135	+12.22890
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	-7.14745	-6.95426	+0.19319	0.11040	+1.74987
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+7.20276	+7.80956	+0.60681	0.06397	+9.48645
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.22782	-0.25990	-0.03208	0.00461	-6.96257
$\mathrm{Re}(A_{0}^{L})\ \gamma$	+9.89863	+10.46061	+0.56198	0.08891	+6.32109
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+1.00000	+1.08477	+0.08477	0.01032	+8.21633
$\mathrm{Im}(A_{00}^{L})\ \alpha$	+1.00000	+1.10417	+0.10417	0.02308	+4.51258
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+1.00000	+1.14789	+0.14789	0.02406	+6.14618
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+1.00000	+1.08179	+0.08179	0.02680	+3.05210

Table 22: Values of the signal, fit mean, difference between the fit mean and signal, standard error and pull mean for an ensemble of 1000 fits using SM signal. Only trained coefficients are included. The lower limit of the $K^{*0}(700)$ mass and the upper limit of the $K^{*0}(700)$ decay width from the 2018 PDG [12] was used for this ensemble. A learning rate of 0.10 was used for this data with β_1, β_2 and ϵ left at their default values.

A.6 SM (0.05 learning rate)

Coefficient	Signal	Mean	Difference	Std. Err	Pull Mean
$\mathrm{Re}(A_{\parallel}^{L})\ \alpha$	-4.17810	-3.63005	+0.54805	0.06600	+8.30354
$\mathrm{Re}(A_{\parallel}^{L})\ \beta$	-0.15184	-0.19032	-0.03847	0.00803	-4.79178
$\mathrm{Re}(A_{\parallel}^{L})\ \gamma$	+6.81832	+5.75909	-1.05924	0.10259	-10.32466
$\mathrm{Im}(A_{\parallel}^{L})\ \alpha$	+0.00859	+0.01408	+0.00549	0.04814	+0.11412
$\mathrm{Im}(A_{\parallel}^{L})\ \beta$	-0.00182	-0.00167	+0.00015	0.00656	+0.02306
$\mathrm{Im}(A_{\parallel}^{L})\ \gamma$	+0.46607	+0.47439	+0.00832	0.07003	+0.11880
$\mathrm{Re}(A_{\parallel}^{R})\ \alpha$	-0.23538	+0.30093	+0.53631	0.07785	+6.88915
$\mathrm{Re}(A_{\parallel}^{R})\ \beta$	-0.00432	-0.10234	-0.09803	0.01192	-8.22550
$\mathrm{Re}(A_{\parallel}^{R})\ \gamma$	+8.00375	+7.46951	-0.53423	0.10689	-4.99788
$\mathrm{Im}(A_{\parallel}^{R})\ \alpha$	+0.16564	+0.09335	-0.07230	0.12363	-0.58476
$\mathrm{Im}(A_{\parallel}^{R})\ \beta$	-0.01310	-0.01725	-0.00416	0.01529	-0.27179
$\mathrm{Im}(A_{\parallel}^{R})\ \gamma$	-0.30668	-0.19488	+0.11180	0.17284	+0.64687
$\mathrm{Re}(A_{\perp}^{L})\ \alpha$	+3.88641	+3.92940	+0.04299	0.04864	+0.88380
$\mathrm{Re}(A_{\perp}^{L})\ \beta$	+0.08527	+0.08899	+0.00373	0.00575	+0.64835
$\mathrm{Re}(A_{\perp}^{L})\ \gamma$	-8.19745	-8.31565	-0.11821	0.08413	-1.40506
$\mathrm{Im}(A_{\perp}^{L})\ \alpha$	-0.09505	-0.23131	-0.13626	0.07858	-1.73397
$\mathrm{Im}(A_{\perp}^{L})\ \beta$	+0.00793	+0.02089	+0.01295	0.00996	+1.30020
$\mathrm{Im}(A_{\perp}^{L})\ \gamma$	-0.07297	+0.19590	+0.26887	0.11953	+2.24944
$\mathrm{Re}(A_{\perp}^{R})\ \alpha$	-0.42358	-0.79915	-0.37557	0.07325	-5.12751
$\mathrm{Re}(A_{\perp}^{R})\ \beta$	+0.02730	+0.13728	+0.10998	0.01115	+9.86296
$\mathrm{Re}(A_{\perp}^{R})\ \gamma$	-7.14745	-6.90653	+0.24093	0.10777	+2.23546
$\mathrm{Re}(A_{0}^{L})\ \alpha$	+7.20276	+7.45098	+0.24823	0.06500	+3.81858
$\mathrm{Re}(A_{0}^{L})\ \beta$	-0.22782	-0.24225	-0.01443	0.00451	-3.19820
$\mathrm{Re}(A_{0}^{L})\ \gamma$	+9.89863	+10.13266	+0.23403	0.09061	+2.58295
$\mathrm{Re}(A_{00}^{L})\ \alpha$	+1.00000	+1.03846	+0.03846	0.01047	+3.67419
$\mathrm{Im}(A_{00}^{L})\ \alpha$	+1.00000	+1.05646	+0.05646	0.02281	+2.47546
$\mathrm{Re}(A_{00}^{R})\ \alpha$	+1.00000	+1.12431	+0.12431	0.02239	+5.55096
$\mathrm{Im}(A_{00}^{R})\ \alpha$	+1.00000	+0.99987	-0.00013	0.02603	-0.00496

Table 23: Values of the signal, fit mean, difference between the fit mean and signal, standard error and pull mean for an ensemble of 1000 fits using SM signal. Only trained coefficients are included. A reduced learning rate of 0.05 was used for this data with β_1, β_2 and ϵ left at their default values.

References

[1] J. Aebischer *et al.*, *B-decay discrepancies after Moriond 2019*, arXiv:1903. 10434 [hep-ex, physics:hep-ph] (2019), arXiv: 1903.10434.

[2] The LHCb collaboration *et al.*, *Test of lepton universality with* $B^0 \to K^{*0}\ell^+\ell^-$ *decays*, Journal of High Energy Physics **2017** (2017) 55.

[3] The LHCb collaboration *et al.*, *Angular analysis of the* $B^0 \to K^{*0}\mu^+\mu^-$ *decay using* $3fb^{-1}$ *of integrated luminosity*, Journal of High Energy Physics **2016** (2016) 104.

[4] A. Khodjamirian, T. Mannel, and N. Offen, *Form factors from light-cone sum rules with B -meson distribution amplitudes*, Physical Review D **75** (2007) 054013.

[5] C. Bobeth, M. Chrzaszcz, D. van Dyk, and J. Virto, *Long-distance effects in* $B \to K^*\ell\ell$ *from analyticity*, The European Physical Journal C **78** (2018) 451.

[6] J. Matias, F. Mescia, M. Ramon, and J. Virto, *Complete anatomy of* $\overline{B_d} \to \overline{K}^{*0}(\to K\pi)\ell^+\ell^-$ *and its angular distribution*, Journal of High Energy Physics **2012** (2012) 104.

[7] S. Descotes-Genon, J. Matias, M. Ramon, and J. Virto, *Implications from clean observables for the binned analysis of* $B \to K^*\mu^+\mu^-$ *at large recoil*, Journal of High Energy Physics **2013** (2013) 48.

[8] U. Egede, M. Patel, and K. A. Petridis, *Method for an unbinned measurement of the* q^2 *dependent decay amplitudes of* $\bar{B}^0 \to \bar{K}^{*0}\mu^+\mu^-$ *decays*, Journal of High Energy Physics **2015** (2015) 84, arXiv: 1504.00574.

[9] R. Aaij *et al.*, *A new algorithm for identifying the flavour of* B_s^0 *mesons at LHCb*, Journal of Instrumentation **11** (2016) P05010.

[10] D. Derkach *et al.*, *Machine-Learning-based global particle-identification algorithms at the LHCb experiment*, Journal of Physics: Conference Series **1085** (2018) 042038.

[11] L. collaboration *et al.*, *Differential branching fraction and angular analysis of the decay* $B^0 \to K^{*0}\mu^+\mu^-$, Journal of High Energy Physics **2013** (2013) 131, arXiv: 1304.6325.

[12] M. Tanabashi *et al.*, *Review of Particle Physics*, Physical Review D **98** (2018) 030001.

[13] J. Beringer *et al.*, *Review of Particle Physics*, Physical Review D **86** (2012) 010001.

[14] L. Hofer and J. Matias, *Exploiting the Symmetries of P and S wave for* $B \to K^*\mu^+\mu^-$, Journal of High Energy Physics **2015** (2015) 104, arXiv: 1502.00920.

[15] *flavio flavour phenomenology in the Standard Model and beyond*, https://flavio.github.io/.

[16] *lejambon/b-decay-unbinned*, https://github.com/lejambon/b-decay-unbinned-machine-fit.

[17] *tf.keras.optimizers.SGD | TensorFlow Core r1.14*, https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/SGD.

[18] *tf.keras.optimizers.RMSprop | TensorFlow Core r1.14*, https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/RMSprop.

[19] *tf.keras.optimizers.Nadam | TensorFlow Core r1.14*, https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Nadam.

[20] *tf.keras.optimizers.Adam | TensorFlow Core r1.14*, https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam.

[21] S. J. Reddi, S. Kale, and S. Kumar, *On the Convergence of Adam and Beyond*, arXiv:1904. 09237 [cs, math, stat] (2019), arXiv: 1904.09237.

[22] *Hyperparameter Tuning with the HParams Dashboard | TensorBoard*, https://www.tensorflow.org/tensorboard/r2/hyperparameter_tuning_with_hparams.

[23] *AutoGraph: Easy control flow for graphs | TensorFlow Core*, https://www.tensorflow.org/guide/autograph.

[24] F. Beaujean, C. Bobeth, and D. van Dyk, *Comprehensive Bayesian analysis of rare (semi)leptonic and radiative B decays*, The European Physical Journal C **74** (2014) 2897.