university of groningen / faculty of science and engineering

Electron Momentum Reconstruction using Supervised Machine Learning at the LHCb

Bachelor Research Project

Name:
Mikuláš Bébr

Student number:
S3070034

Abstract

This research focuses on improving the momentum reconstruction method employed at the LHCb by using supervised machine learning. An overview of the principle of lepton universality violation and its implication is provided. A performance comparison between the Decision Tree regression algorithm and Polynomial regression using the least-squares method is given as well as the drawbacks of each method. The main result is the plot of the residual sum of the reconstructed momentum and the true value of the momentum from simulated data of the LHCb. The research concludes that although the supervised learning machine algorithms do provide a general decrease in the standard deviation of the residual sum, it does inflate the distribution around 0, effectively reducing the sharpness of the peak of the distribution.

First Examiner:
dr. M.C. van Veghel
Second Examiner:
dr. ir. C.J.G. Onderwater

July 8, 2022

Contents

Acknowledgements

A very special thanks to Maarten van Veghel for his patience and sharing his expertise throughout this research project, Kristof de Bruyn and Gerco Onderwater for their helpful insights during the meetings, as well as all the fellow Bachelor students in the LHCb group for sharing views and observations.

1 Introduction

Modern science, in the general sense of the word, is a self-correcting process [1]. Testing hypotheses that would potentially adjust or completely alter older conventions are at the core of the scientific method, which effectively ensures that scientific beliefs stay up to date, and in line with current experimental results. One striking example of this phenomenon in Physics is the Standard Model (SM). Developed in the second half of the 20th century, and building on the recent advancements in the revolutionary Quantum Field theory, the SM provided (and continues to do so) an extraordinarily robust theory of the elementary constituents of our universe, and their interactions. The SM served to successfully predict countless physical mechanisms that were subsequently experimentally confirmed. One of the latest, most prominent results yielded by the SM is the correct prediction of the Higgs boson, whose existence has been experimentally determined in 2012, when the theorized Higgs boson has been discovered in the Large Hadron Collider (LHC) in CERN, Switzerland. Following this breakthrough in the world of particle physics, the SM has been adjusted accordingly and incorporated the Higgs boson and its interactions. As powerful as the SM is, it still too has many considerable drawbacks that keep the particle and theoretical physicists busy: the lack of gravitational force in the model and its inability to incorporate dark matter and dark energy, to name a few. These limitations result in the fact that the SM accounts for only 5% (!) of the energy-matter content of the universe [2]. It is therefore only logical that there is immense focus on exploiting those cracks in the SM, to provide more accurate descriptions of the universe. As Bifani et. al. puts it: "A more global theory that extends the SM at higher energies and shorter distances could provide an answer to some of these questions, which are at the core of modern particle physics." [3]. This project will propose a technique that will improve the reconstruction of electron momentum in the LHCb experiment by answering the question

To what extent does the use of machine learning improve the electron momentum reconstruction performance of the LHCb?

1.1 Physics of the Standard Model

The current version of the SM provides a description of the elementary particles and their interactions in the language of Quantum Field theory. In this framework, particles are the results of excitations of fields. Those particles are called fermions and have 1/2 spin, a visual classification is shown in figure 1. The model also includes 3 elementary forces, the strong force, mediated by the massless gluon gauge boson, the electromagnetic force, mediated by the massless photon, and the weak force which is mediated by the charged massive $W^{\pm}$ and neutral massive Z bosons. Notice that gravity is omitted, sometimes postulated by a hypothetical force mediator called the graviton. Fermions can be further divided into 6 quarks and 6 leptons, both consisting of 3 families (generations) with increasing mass. Quarks interact via the strong force and constitute hadronic matter, such as protons neutrons or the B-meson that will become more important later in this paper. All fermions interact via the weak force, which in the case of quarks leads to the allowed violation of

flavour [4].

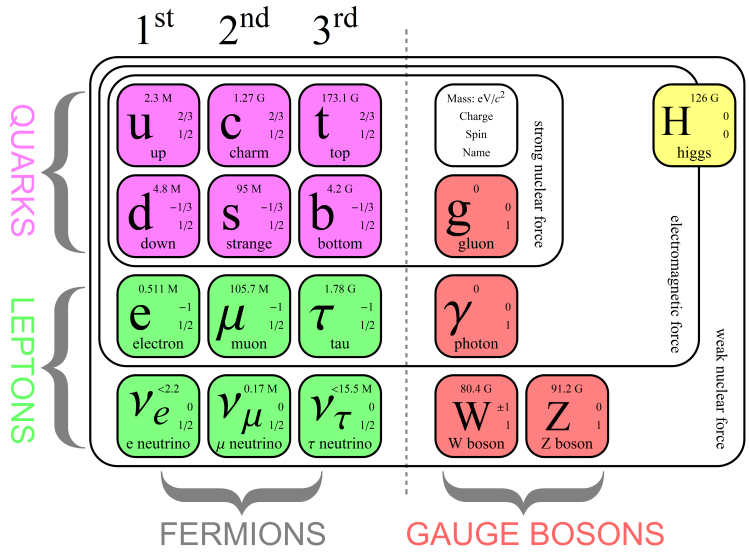


Figure 1: Categorization of elementary sub-atomic particles according to the SM [5]

1.2 Lepton Universality

According to the SM, the 3 leptonic doublets are identical, apart from their differing mass, increasing from left to right in figure 1. Their interaction with other particles by means of the electroweak forces referred to as coupling to the $W^{\pm}$ and Z bosons in the language of quantum field theory, is therefore supposedly also identical. Any deviation from this accidental property, called Lepton Universality (LU or LFU for Lepton Flavour Universality) inherent to the SM, would imply new physics beyond the SM. Hence, it is the perfect testing ground for many recent experiments, the LHCb among others. [3]. Testing LU has proven to be in accordance with SM predictions for electroweak decays of the $W^{\pm}$ and Z bosons for the first two lepton generations. However, the branching fraction ratio of the W-boson decaying into 3rd generation leptons and 1st or 2nd generation leptons shows tension with SM predictions at the level of 2.6 σ [3].
Proton-proton collisions provided by the LHC provide an unprecedented source of B-mesons (containing the b quark such as B^+ $(u\bar{b})$ or B^0 $(d\bar{b})$), whose rare decays are used to test the LU. Even though the more clinical creation of $b\bar{b}$ quark pairs in other experimental setups, called b-factories, yield less noise, courtesy of the orders of magnitude higher yield of $b\bar{b}$ pairs in the LHCb it provides a great setting to examine the rare decays, provided effective noise suppression [4]. The quark transitions of the beauty quark used to probe LU in the LHCb can be divided into two categories:

$$b \to c\ell^-\bar{\nu}_\ell \tag{1}$$

and the much rarer

$$b \to s\ell^-\ell^+ \tag{2}$$

where $\ell^{\pm}$ represents any of the 3 flavours of (anti)leptons and $\bar{\nu}_{\ell}$ its corresponding antineutrino. In the rest of the thesis, we will focus exclusively on the latter transition, which emerges in the rare decay mode of the dataset studied in section 3, namely

$$B^{+} \to K^{+} J/\psi (\to e^{+} e^{-}). \tag{3}$$

This decay channel via Flavour Changing Neutral Current (FCNC) above is forbidden at tree level in the SM, which leads to a high suppression of the channel. This is ideal for the study of LU violation, as the events are not drowned in signal [6].
The branching fraction ratio of the above channel with the branching involving muons R_{K^+} is then used to determine the deviation of the results from the LU predicted by the standard model. Although other observables can be analysed, such as the difference in angular properties between the two channels, this method is more accurate because the ratio cancels out the so-called "hadronic uncertainties" shared by both branching fractions, yielding the result more accurate [3].
The resulting ratio of branching fractions in the above decay mode measured by the LHCb is presented in figure 2.

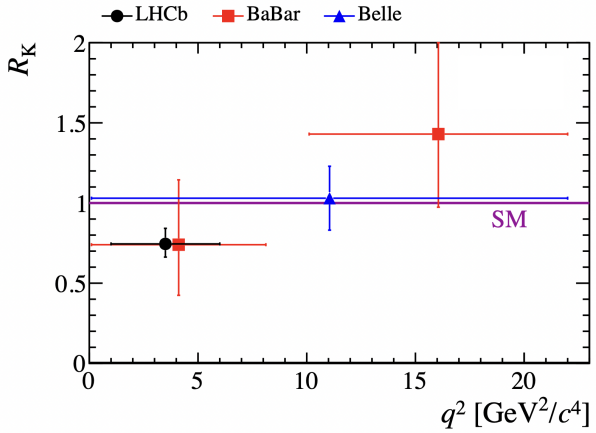


Figure 2: Results of the ratio R_K of branching fractions $B^{+} \to K^{+}\mu^{+}\mu^{-}$ and $B^{+} \to K^{+}e^{+}e^{-}$ in the LHCb [7] compared to other experiments (BaBar [8], Belle [9])

Note that to cancel out experimental uncertainties, a double-ratio is used, normalising the non-resonating mode (excluding the intermediate J/ψ particle) to the mode in eqn. (3). The resulting ratio R_K is $0.745^{+0.090}_{-0.074} \pm 0.036$ where the first uncertainty corresponds to statistical and the second to experimental errors. The measurement is taken over a range of q^2 s.t. $1 < q^2 < 6\ GeV^2/c^4$, which corresponds to the difference between the momentum of the B^{+} and K^{+}, the invariant mass squared of the di-lepton pair [7]. The results show a deviation below the predicted value by the SM by 2.6 σ. Other decay channels containing *e.g.* the transition in eqn. 1 and their respective ratios of branching fractions show even more significant results, deviating up to 4 σ from the SM predicted value. None of these results is, however, significant enough to provide clear evidence of New Physics, as a threshold of 5σ is generally required in particle physics [4].
To achieve the deviation of 5 σ, and consequently confirm unambiguous evidence of New

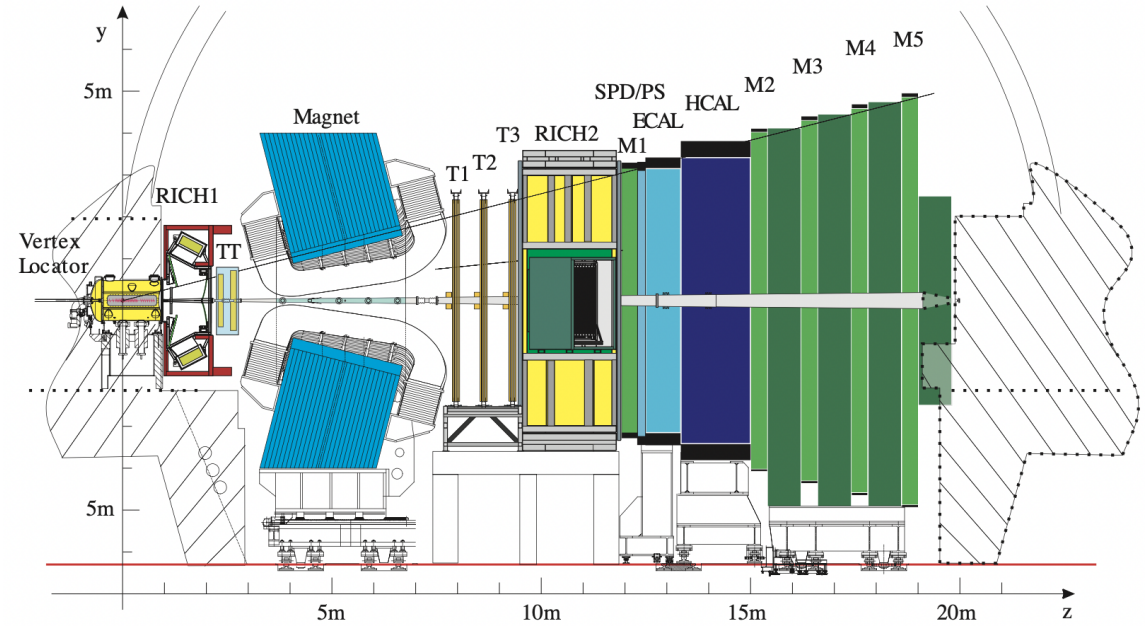


Figure 3: Schematic cross-section of the LHCb arm along the y-z plane [10]

Physics, more data has to be analyzed and the resolution of the detectors has to be improved. The rest of the paper will be devoted to improving the detection resolution by employing Machine Learning computational methods in the reconstruction of electron momentum in the LHCb detector.

2 LHCb

The Large Hadron Collider beauty (LHCb) is one of the 4 main experiments conducted in the complex of the Large Hadron Collider (LHC) in CERN, Geneva. It utilizes a single-arm spectrometer design with angular coverage from 10 mrad to 300 mrad and 250 mrad in the horizontal bending and vertical non-bending plane, respectively.

Besides the decay channels mentioned in section 1.2, the LHCb probes also other decay channels to prove LU violation, as well as a range of other research topics such as lepton flavour violation, CP violations, etc., that point toward the discovery of new physics beyond the SM. All of these experiments require accurate detection, path reconstruction and identification of the secondary particles resulting from the pp collision. This is why the LHCb employs a range of sub-detectors along the particle path. Only the sub-detectors relevant for this project are discussed here. A detailed description of all of the hardware components, as well as the software employed for data collection and processing, can be found in [10] and [11], respectively.

The LHCb detector employs a range of various sub-detectors, each providing measurements of different parameters of the incoming particles. The physical location of each detector

can be seen in Figure 3. Note that the bending plane of the magnet is horizontal (x-z). The detectors can be categorized depending on their relative location to the magnet. Those sub-detectors that the particle encounters prior to the interaction with the magnetic field of the magnet are called the 'upstream' detectors, while those encountered after the magnet are called 'downstream' detectors. This distinction is important, as it distinguishes between energy losses that are furthermore affected by the magnet and those that do not get deflected.

2.1 Vertex Locator (VELO)

The first detector is located in the immediate vicinity of the interaction region. It is therefore used to precisely measure the coordinates of the particles close to the interaction region. It consists of a series of silicon layers with alternating segmentation to provide r and ϕ [10] measurements. The choice of the polar coordinates is justified by its computational efficiency over its Euclidian counterpart [6]. The VELO measures the position of secondary vertices, a characteristic behaviour of the b and c- hadron decay [12]. The VELO detector is retractable to protect it from harmful radiation. It operates in its vacuum, separated from the rest of the detector by a thin aluminium sheet. This aluminium sheet is an important source of upstream energy loss by means of Bremsstrahlung.

2.2 Trackers

The Tracker Turicensis (TT) is a silicon tracker used to locate the travelling particle in the x-y plane with a resolution of approximately 50 μm [10]. The Inner Tracker (IT) is similar to the TT, but is located downstream of the magnet and spans a smaller area.
The Outer Tracker located downstream of the magnet consists of 3 tracker stations, each comprising several straw-tube modules, and serves to track charged particles and measure their momenta with relative momentum resolution of $\frac{\Delta p}{p} \approx 0.4$ [10]. This allows for very precise reconstruction of the B-hadron invariant mass. All trackers are locations of high Bremsstrahlung momentum losses, however, only the TT is located upstream, and is relevant for this research.

2.3 Calorimeter setup

The Scintillator Pad Detector (SPD) is a scintillator placed in front of the calorimeters. Using photomultiplier tubes, it serves to distinguish between charged and neutral particles. Photons and electrons will interact with the lead slab placed further downstream between the SPD and Pre Shower detector (PS), and initiate an electromagnetic shower. The latter is then detected using the PS scintillator. The secondary particles then interact further downstream with either the Electromagnetic CALorimeter or the Hadronic CALorimeter. Data from this set of sub-detectors are mostly used for particle identification, by probing their signature showers.

2.4 Magnet

In the centre of the LHCb detector is located the dipole magnet with a highly non-homogenous magnetic field of approximately 4 Tm. It is used to deflect charged particles

and based on their curvature measure their momentum.

2.5 Electron in LHCb

The electron serves as a valuable secondary particle and by analyzing its properties such as momentum and transverse energy, one can infer the properties of the system in question. As mentioned above in section 1.2, the identification and momenta of the electron-positron pair in $B^+ \to K^+ J/\psi(\to e^+e^-)$ are used in the determination of the branching fraction ratio. Due to its low mass, however, the electron interacts heavily with matter that it is exposed to. This behaviour is unfavourable as these interactions alter the otherwise straight path of the electron (besides the intentional and controlled magnet bending), but more importantly, the electron loses its energy. The most dominant mechanism of electron energy loss is Bremsstrahlung, which occurs when a charged particle travels through the electric field of the nucleus of a medium.
In the LHCb, such energy loss can occur either upstream or downstream of the magnet. In the latter case, the photon carrying the energy lost by the electron continues in the approximately same direction as its parent electron. In the former case, however, the electron is deflected by the magnet in the bending plane, but the Bremsstrahlung photon carries on in the same direction unaffected by the magnet. The photon and electron will therefore arrive at the calorimeter at radically different positions, posing a challenge for the reconstruction of the initial electron energy. This behaviour is illustrated in figure 4. The LHCb employs a clustering algorithm that attempts to link energy deposits in

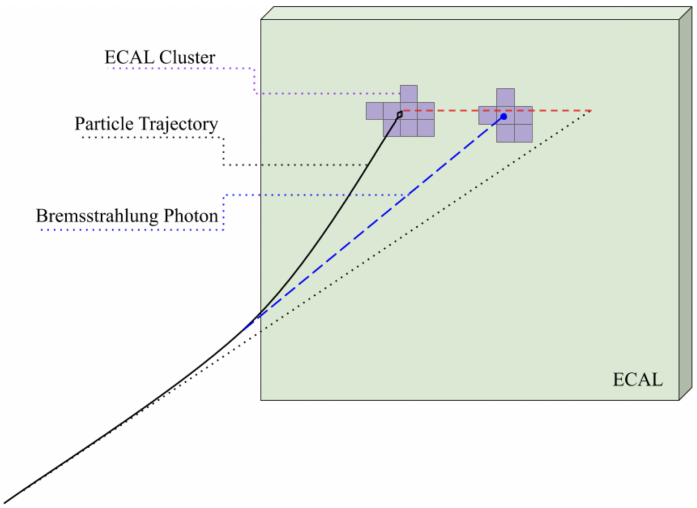


Figure 4: Schematic of a charged particle detection in the ECAL, bending in the magnet after upstream energy loss via Bremsstrahlung photon emission. Figure from [6]

the calorimeter that could correspond to Bremsstrahlung energy losses to their respective electrons.

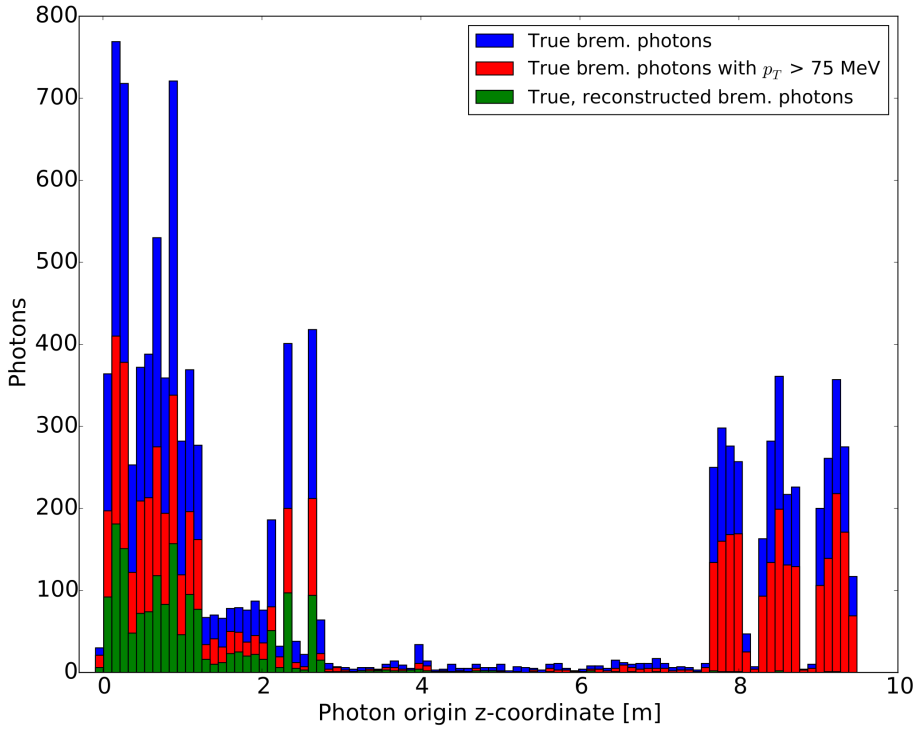


Figure 5: Photon origin z-coordinates. Blue: generated brem. photons, red: generated brem. photons with transverse momentum greater than 75 MeV, green: generated brem. photons which are reconstructed. Figure from [14]

2.6 Bremsstrahlung

As mentioned above, the main energy loss of a relativistic electron happens through Bremsstrahlung (abbreviated as Brem in figures and labels). This process happens when a charged particle is accelerated in the electric field of a nucleus of the medium it travels through, and as a result of this acceleration, a photon is emitted. For relativistic charged particles ($E_0 > mc^2$) the cross-section, σ_{brem}, of this interaction can be approximated by

$$\sigma_{brem} \approx \alpha Z^2 (\frac{e}{mc^2})^2 \tag{4}$$

where α is a proportionality factor, Z the nuclear charge and m the rest mass of the incoming particle. From eqn. 4 it is apparent that due to the inverse squared mass-dependence, the electron, the lightest charged lepton, is affected more by Bremsstrahlung than other particles [13].

In figure 5 we can see the locations at which most of the Bremsstrahlung occurs. Note that there are multiple threshold conditions that have to be met for the reconstruction algorithm to even consider the photon for reconstruction. Namely, the photons have to have total momentum $p > 100$ MeV, and transverse momentum $p_T > 75$ MeV. From figure 5 we can see that about 50% of photons with $p > 100$ MeV have $p_T > 75$ MeV. Of these, again about 50 % are successfully reconstructed. This leads to a relatively low amount of data available

for momentum reconstruction. This will be further elaborated on in Section 3.4. Note that none of the photons emitted after the magnet is reconstructed. This is because, as mentioned earlier, if the Bremsstrahlung energy loss occurs in layers downstream of the magnet, the emitted photons are detected in the same place as their parent electron. [14] Photons yielding from such high-energy electrons keep the same direction as the electron.

3 Electron Energy Reconstruction

As was introduced in section 1.2, improving the electron detection resolution can contribute to a more accurate determination of the branching fractions of decay channels involving e^+e^- pairs. The goal of this section is to propose a more accurate way of reconstructing energy lost by the electron while travelling through the detector using a supervised Machine Learning technique. In this section, the dataset, the theory behind the chosen ML algorithm as well as the algorithm performance are discussed.

3.1 Simulated data

For this research project, we consider the simulated data of the $B^+ \to K^+ J/\psi(\to e^+e^-)$ resonating decay channel. There are two separate applications that serve to simulate the LHCb data, called Gauss and Boole. Gauss is used for particle generation and the physical behaviour of the particles in all of the subdetector layers of the spectrometer. Boole is an emulation of the detector response and the digitization process. Once the data is digitized, it follows the same path as real raw data acquired by the detector for processing. This method is a very reliable way of generating data with known values as well as reflecting the behaviour and various resolution limitations of both the software and hardware employed [15].
The provided dataset includes about 19×10^6 events with 42 parameters. These include unprocessed data, such as the various Bremsstrahlung correction variables, as well as processed variables, such as the particle identification or its momentum.

For the purpose of electron momentum reconstruction we are mostly interested in the following variables: p_{true}, which denotes the original momentum of the electron (provdied by the simulations), p_{e^-} the electron's measured momentum by the detector excluding Bremsstrahlung momentum reconstruction of any kind, p_{brem} the sum of all Bremsstrahlung photon momenta recovered by the detector, corresponding to a given cluster linked to an electron detection event, and finally, p_{recon} which represents the reconstructed momentum of the electron, based on the given reconstruction method. To filter (mask) relevant data corresponding to electrons in the required energy range, particle identification values are used as well as parameters describing the electron track behaviour in the detector such as its presence in the ECAL. The distribution of p_{e^-}, p_{brem} and p_{true} is plotted in figure 6.

The dataset has been filtered using bitwise masking in Python, to select events that correspond to electrons (`'Track_MC_electron_signal'==1`), and tracks that travel through the entirety of the detector and are detected in the Ecal (`'Track_Type'==3` and `'Track_InEcal'==1`, respectively). Furthermore, a momentum threshold has been set at 1 GeV/c, due to poor reconstruction capabilities below this threshold.

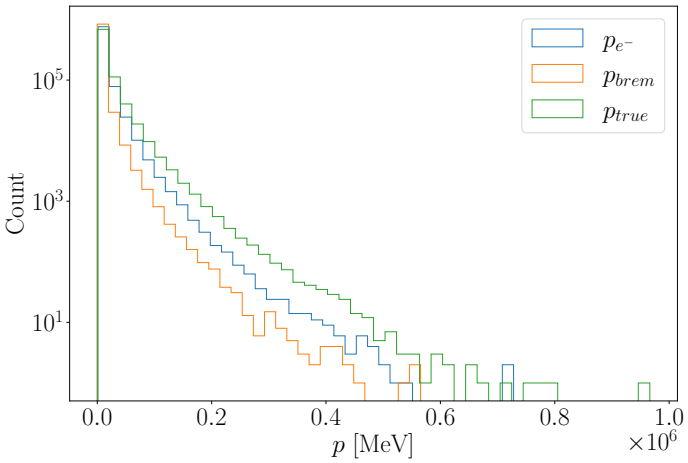


Figure 6: Momentum distribution of electrons in $B^+ \to K^+ J/\psi(\to e^+e^-)$ in LHCb (log y-scale, bin=48)

3.2 Minimisation of $p_{recon} - p_{true}$ residual using supervised Machine Learning

Because the dataset is a result of a simulation, we can exploit the fact that the true value of any parameter is known, as well as its measured counterpart. It is therefore possible to employ a supervised machine learning algorithm that uses the *true* data points to learn the regression to fit and predict data points based on measurements of their parameters. A supervised machine learning (ML) algorithm uses a sub-sample of the dataset, called the training set, to fit the data points and the rest of the dataset to test the quality of the fit. The dataset usually contains multiple independent variables (features in the language of ML) and can contain multiple dependent variables that the algorithm aims to predict. The training sample is usually about 3/4 of the whole dataset, however, this can be altered for specific needs of a given algorithm–some algorithms are more prone to the so-called "data overfitting", in which case the algorithm fits the training set too *tightly*, taking into account the outliers rather than the general trend of the training dataset. This results in poor performance when making predictions on the testing set. To determine the performance of the fit, a scoring method can be employed on the testing sample. The number of total events in our dataset (after masking and the lower energy bound) is approximately 884000, which after train/test -splitting yields 707154 training samples and 176789 testing samples. The choice of algorithms has been made by inspecting their *out-of-the-box* performance. This approach has been taken to allow analysis and fine-tuning of the performance of two algorithms, rather than comparing the whole range of available ML methods. This point will be further elaborated on in the discussion section.

3.2.1 Polynomial regression

The polynomial regression (PR) algorithm is based on the least-squares fitting algorithm. The least-squares algorithm fits a vector $\vec{w}$ such that the residual sum of squares between the weighted data points and the true value is minimized. In mathematical terms, this

corresponds to approximating the overdetermined system

$$\min_{\vec{w}} ||X\vec{w} - \vec{y}|| \tag{5}$$

where the matrix X represents the dataset, $\vec{w}$ the weight vector and $\vec{y}$ the target variable [16]. to achieve a polynomial regression, one can make use of the least-squares algorithm by pre-processing the dataset. By cross-multiplying features, polynomial terms up to desired order r will be achieved. In order to allow for lower powers of a given feature than r, a unity feature is added. In practice, if a dataset contains $n = 3$ features x_1,x_2 and x_3, and an order 2 polynomial in the regression is required, the feature set is transformed to 1×1, $x_1 \times 1$, $x_1 \times x_2$, ... , x_3^2. Note that the first trivial feature corresponds to the intercept–a constant term of the regression. The number of features is now given by the binomial coefficient ${}^{(n+r-1)}C_r$, where n is the original number of features, and r is the desired order of the polynomial.
The goal of this project is to improve on the current method used by the LHCb, which can be thought of as a 0th order regression—addition of the recovered photons momentum p_{brem} to the measured p_{e^-} momentum with a coefficient 1. Besides its default performance, this algorithm has been chosen due to its simplicity of interpretation of the results. The vector $\vec{w}$ provides a clear weighting coefficient for each feature, which directly reflects its importance in the regression. This advantage is, however, reduced with the introduction of higher order terms. Additionally, this method is remarkably fast, albeit the cross multiplication of features to obtain polynomial terms consumes exponentially more time with each higher order.
The polynomial order r has been chosen by hand to maximise the score of the algorithm upon testing to be 2.

3.2.2 Decision Tree Algorithm

The second supervised machine learning algorithm of choice is the decision tree (DT). The decision tree process can be described as follows: given a learning sample, the algorithm divides the data points according to a condition. This condition is chosen to minimize standard deviation within the given categories (a measure of impurity). This process is repeated, each time branching the given category into two until the required tree depth (maximum number of branches) is met. Note that the algorithm can be used for both discrete categorization problems, as well as for continuous distribution problems. In both cases, however, the resulting fit is a categorization with an arbitrary number of categories. The number of categories is given by the leaves (endings of branches that do not divide any further) and is directly related to the tree depth. The prediction of the desired value is based on an average of data points within the category it. In certain situations this algorithm can be prone to overfitting the data to the training sample, making it non-optimal for the testing sample. In that case, one can additionally use a threshold for a minimum number of data points at each leaf. This is another condition that can terminate the branching of the tree. Similarly to the polynomial regression, the choice of the decision tree to reconstruct the electron momentum has two arguments, in addition to its default performance. Firstly, it is much faster to process than other algorithms, such as a neural network, where the latter on some occasions took over 10 times as long for the same dataset. Furthermore, the *decisions*

that the algorithm makes to fit the training data are easier to interpret and visualise, and the weight of each parameter in the algorithm can be inferred by examining its relative frequency in the decision-making, and its corresponding impurity. Additionally, the design of the algorithm is radically different to the least-squares method. Comparing them can provide interesting insight into their behaviour and success in fitting the same set of data.

To adapt the algorithm to the needs of the reconstruction task and to prevent overfitting, the `min_samples_leaf` parameter has been set to 100, meaning that each leaf must contain at least 100 data points. The `max_depth` of the tree has been set to 13. These parameters have been set by hand, by monitoring the performance of the algorithm while varying them.

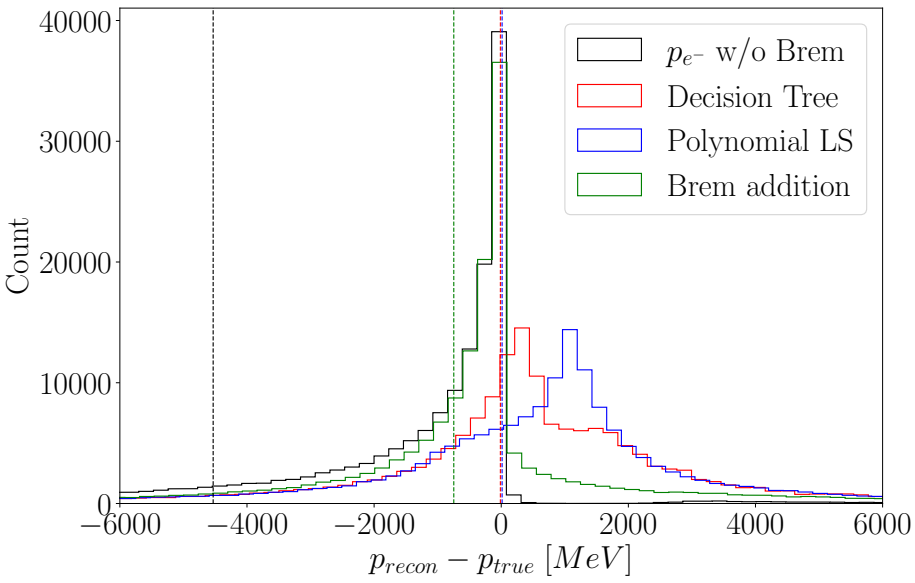
3.3 Results

Figure 7a shows the residual sum of the measured and true momenta $p_{recon} - p_{true}$ for the predictions made by the Decision Tree regression, polynomial least-squares regression, and the simple Bremsstrahlung addition given by $p_{recon} = p_{e^-} + p_{brem}$. A statistical analysis of the distributions is given in Table 1, along with various parameters describing the algorithm performance per reconstruction method. Figure 7b shows the same plot on a logarithmic y-scale, and -10× zoomed-out x-scale to illustrate the behaviour of the distributions at their tails.

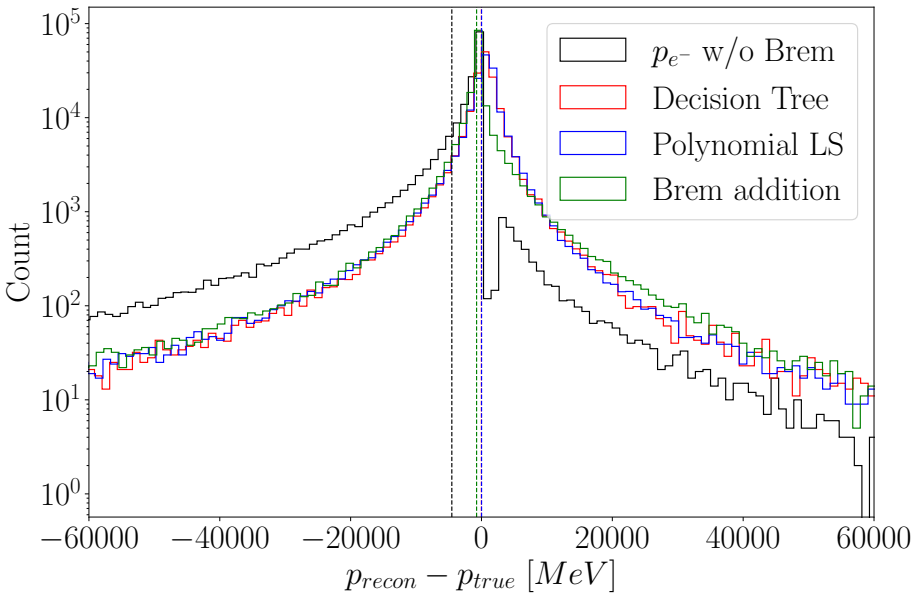
Regression	Mean μ	SD	MAE	Score	$W_{99.8}$	W_{95}	W_{50}	Median
Nobrem	-4532.2	13841.2	5157.2	-	192893.7	39605.9	3602.2	-894.5
BremAdd	-746.5	11435.7	3679.4	-	196739.7	30557.9	1362.8	-237.3
Poly Regr	3.4	10303.3	3709.3	0.8312	177878.4	26495.5	2549.5	808.4
DT	-3.3	10410.0	3548.5	0.8277	182254.6	25994.7	2490.8	369.4

Table 1: Decision tree performance analysis. All columns are in [MeV] units, except score which is a fraction

For both the Polynomial regression and the decision tree, the distribution average has shifted from $\mu = 746.5$ MeV towards 3.4 Mev and -3.3 MeV, respectively. This behaviour is expected, as the algorithms aim to minimize the standard deviation (average of the residual sum of squares), which corresponds to the mean at 0, physically representing an average momentum reconstruction close to the true value.
Figure 8 shows quantile lines. A quantile represents the value below which a fraction of the data points lies. The windows W_{50}, W_{95} and $W_{99.8}$ are then simply the inter-quantile range for 0.25 and 0.75, 0.975 and 0.025 and 0.999 and 0.001, respectively.
The polynomial regression yields approximately a 10% decrease in SD and about a 1% increase in Mean Average error over the simple Bremsstrahlung addition. The window $W_{95}(Poly)$ containing 95% of the data has decreased by 13% and the window $W_{50}(Poly)$ containing 50% of the data has almost doubled.
On the other hand, the DT yields approximately a 9% decrease in SD and about 3.5% decrease in Mean Average error over the simple Bremsstrahlung addition. The window $W_{95}(DT)$ containing 95% of the data has decreased by 15% and the window $W_{50}(DT)$ con-



(a)



(b)

Figure 7: Distribution of the residual sum of measured (reconstructed) and simulated electron momentum $p_{recon} - p_{true}$ [MeV] using DT and PR, 0th order regression and raw p_{e^-} data without Bremsstrahlung added. Vertical lines represent average values of distribution by colour, (zoomed in for clarity, bin=5000 over full range) (a), logarithmic y-scale, increased range (bin=500 over full range) (b)

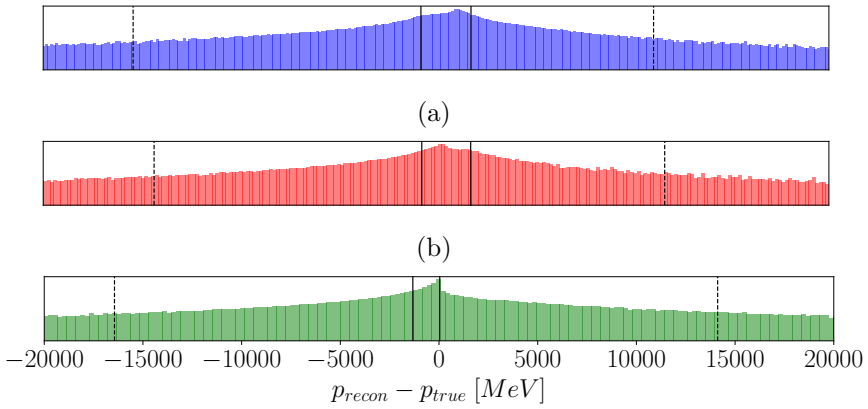
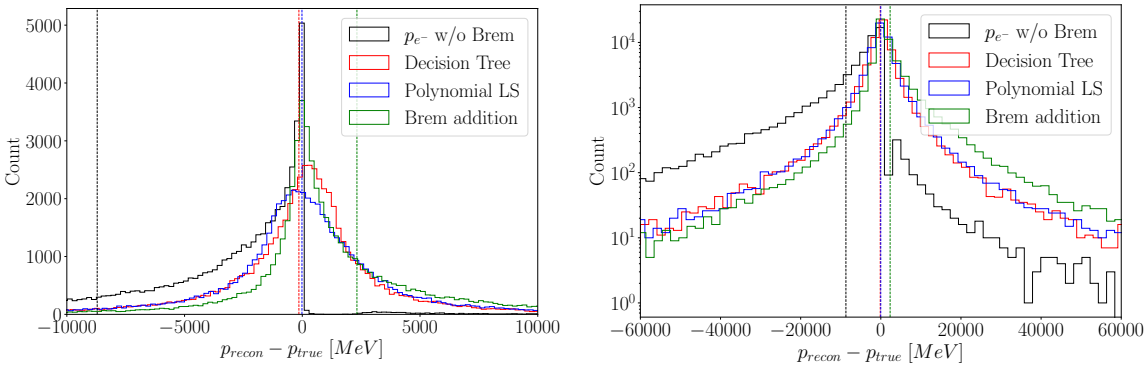


Figure 8: Quantile lines W_{95} (dashed) and W_{50} (solid), log y-axis scaled to fit distribution. PR (a), DT (b), $p_{recon} = p_{e^-} + p_{brem}$ (c)

taining 50% of the data has increased by 80%.
For both regression methods, this implies that the tails of the distribution have narrowed down, but the central part of the distribution has been spread out by the fit. This is in accordance with the values of the Mean Average error and the Standard deviation. The former considers residual sums to the 1st power, whilst the latter uses squared errors—outliers (data points lying on the tails of the distribution) which are more severely penalised in the SD than the central data points. Although the MAE stays approximately constant for both fitting methods, the SD decreases.
In terms of differences between the two algorithms, the most noticeable is the peak shift. According to Figure 7a, the peaks of the two distributions do not align with the mean. Both peaks are shifted in the positive x-direction, but the Polynomial regression is far more offset than the DT. Although the mean of the distribution is counter-balanced by the outliers on the opposite tail and coincides with $\mu \approx 0$, those peaks represent the bins with the highest counts, hence the most frequent estimations are offset by about 250 MeV for the decision tree, and about 1000 MeV for the Polynomial regression. From Figure 7b it can be seen that away from the origin both tails (red and blue) behave very similarly.

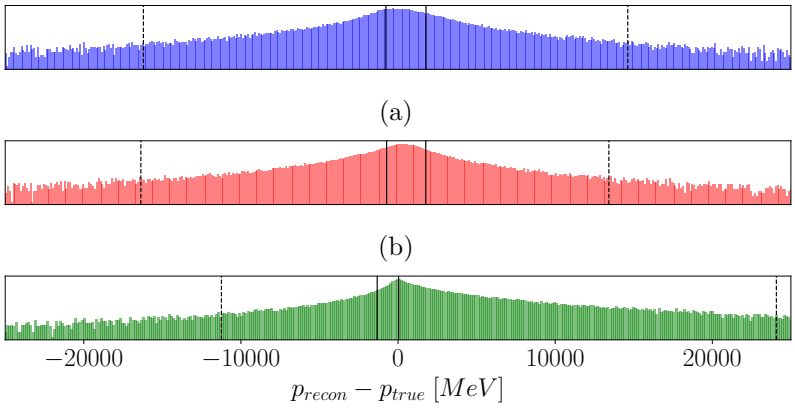
3.4 Masking $p_{brem} = 0$ events

As mentioned in 2.6, only a very small portion of Bremsstrahlung photons do get recovered and associated with the parent electron by clustering. In fact, for the dataset used in this research, there are almost 884 000 events corresponding to electron detection of which about 582 000 (66%) have $p_{brem} = 0$. This clearly cannot reflect the underlying physics correctly, as we have shown that for electrons the Bremsstrahlung cross-section is very large due to its inverse mass-squared proportionality. The probability of an electron-emitting 0 Bremsstrahlung photons is therefore extremely unlikely. A solution to this is filtering the dataset such that only events with $p_{brem} \neq 0$ are considered. Note that this decreases the sample size by 66%. The results of the filtered data are shown in Figures 9a and 9b.



(a) (b)

Figure 9: Distribution of the residual sum of measured (reconstructed) and simulated electron momentum $p_{recon} - p_{true}$ [MeV] using DT and PR, 0th order regression and raw p_{e^-} data without Bremsstrahlung added. Data points with $p_{brem} = 0$ filtered out. Vertical lines represent average values of distribution by colour, (zoomed in for clarity, bin=3000 over full range) (a), logarithmic y-scale, increased range (bin=300 over full range) (b)



(a) (b) (c)

Figure 10: Quantile lines W_{95} (dashed) and W_{50} (solid), log y-axis scaled to fit the distribution. Data points with $p_{brem} = 0$ filtered out. PR (a), DT (b), $p_{recon} = p_{e^-} + p_{brem}$ (c)

The SD of the distribution has now decreased by almost 15% from 11637 MeV for the Bremsstrahlung addition to 9925 and 10020 MeV for PR and DT, respectively. Also, the scoring of both algorithms is 0.88 (PR) and 0.87 (DT), which is almost a 6% increase over the unfiltered dataset in Section 3.3. A very important difference with this filtered dataset is that the distribution of the residual sum $p_{recon} - p_{true}$ is now skewed to the positive direction, *i.e.* the average value of the residual sum is 2320 MeV, as opposed to the -746 MeV for the unfiltered dataset. This is a very surprising result, which will be discussed in the following section.
In terms of the % windows, the behaviour is comparable to the dataset in Section 3.3. The action of decreasing the SD does decrease the size of the 95% window, but widens the 50%

window, as shown in Figure 10.

4 Discussion

4.1 Low Bremsstrahlung photons recovery rate

Probably the most important point of discussion is the lack of Bremsstrahlung photons in general. After masking of the data, as described in section 3.1, about 884 000 data points are corresponding to electron deposit events. From these, however, 582 000 do not have any Bremsstrahlung energy associated. That means that almost 66% of the electron detections are not paired with any Bremsstrahlung photon cluster. A solution to this issue is to simply only consider events in which some Bremsstrahlung photons are detected. This, however, reduces the dataset by 66%. It remains for future research to determine whether this percentage of 0 clustered Bremsstrahlung photons per electron detection is specific to the dataset in question, or inherent to all measurements taken at LHCb. To prevent the drastic decrease in the sample size, one could mark those events where $p_{brem} \neq 0$, and treat them separately: use it as a training set for the supervised ML. On the other hand, events for which $p_{brem} = 0$ do not need to be included in the reconstruction process, as they do not have any data to reconstruct from, other than a constant term that would shift the mean of $p_{recon} - p_{true}$ distribution towards 0.
When this suggested filtering has been applied in Section 3.4, and data points (events) where Bremsstrahlung photons with momentum $p_{brem} = 0$ are excluded, besides indeed improving the performance of both ML algorithms, another interesting and surprising change takes place. The distribution of $p_{recon} - p_{true}$ using the Bremsstrahlung addition now has an average above zero. This means, that on average, this method tends to over-estimate the true momentum of the electron. This is very surprising, as it shows that the unreasonably large number of events with $p_{brem} = 0$ in the unfiltered dataset compensates for the "over-clustering" tendencies of the algorithm that determines p_{brem}. In other words, *when* $p_{brem} \neq 0$, it significantly over-estimates the actual momentum of the electron, presumably by including background noise into the cluster corresponding to the given electron. This behaviour should be further examined, and an explanation for the excessive momentum estimation should be provided. One possible explanation is the lack of final-state radiation in the value of p_{true}. This, however, is not likely to be responsible for the shift of the whole spectrum. Also note that if a constant value corresponding to the final state radiation would be added to p_{true}, this would shift the whole distribution toward the negative side together with the peak, which is now correctly positioned at 0.

4.2 Sharpness-compactness trade-off

Another observation worth emphasizing is that there seems to be a general trade-off between the sharpness of the data distribution in the central region of the distribution, versus its compactness at the ends, as illustrated in Figure 11.

This can be quantitatively demonstrated in two ways: Firstly, from Table 1, when the 0th order regression is used for the momentum reconstruction, the standard deviation of the distribution of the residual sum $p_{recon} = p_{e^-} + p_{brem}$ is higher than for either supervised

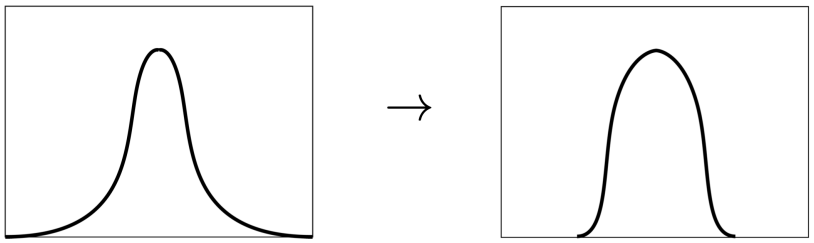


Figure 11: Qualitative demonstration of sharpness-compactness trade-off. Decrease in SD from left to right

ML learning methods. This narrowing of the distribution at its tails (decreasing the SD due to its aforementioned higher sensitivity to outliers) is counteracted by an expansion of the distribution in the central region. This can be seen by examining the change in W_{95} and W_{50}. While the former decreases, the latter increases. A similar behaviour, although much less pronounced, can be seen between the two ML algorithms. Whilst the PR is an improvement in SD over the DT, one has to consider 99.8% of the data to notice the lower spread ($W_{99.8}(DT) > W_{99.8}(PR)$). All other central windows (W_{95} and W_{50}) yield higher results for the PR than the DT. This implies that although in the general, the distribution as a whole is more compact for the PR than DT (lower SD), this is only thanks to the 0.2% of data lying on the extremities of the distributions. One can conclude that the higher the decrease in SD by means of a reconstruction technique, the higher the spread out in the central region.

4.3 Choice of algorithms and their performance

Furthermore, the justification for the choice of reconstruction algorithm will be discussed. As mentioned in section 3.2, the DT and PR algorithms were chosen a) by inspection of their *out-of-the-box* (using default settings) performance, such as scoring, standard deviation and mean shift. The reasoning behind this choice is that optimising a large number of algorithms with a wide range of parameters is very time-consuming (computationally challenging), and the emphasis in this research project is therefore on the optimisation of two such algorithms. Evidently, including more algorithms in the selection could yield better results, however, the scores differences are marginal (disregarding the elapsed time, and hence computational efficiency). This points toward the question, of whether there exists a limit to the performance of all supervised ML techniques for a given dataset. If so, it may be useful to introduce another scoring system which would reflect not the absolute performance of the regression on the test dataset, but relative to this limit that can be achieved. The existence of such a limit is logical, as fitting perfectly the training data results in overfitting and poor performance on the testing data, but a *loose* fit on the testing data may neglect the more subtle behaviour of the studied dataset. The performance limit should be found therefore somewhere between these two cases. It has been pointed out, that the Neyman Pearson lemma treats this idea, it is however out of my abilities to present any of the encompassing theories.
Another suggestion for improvement of the fitting abilities of all supervised ML algorithms is simply more training data. This can be achieved by training on datasets including

electrons corresponding to different decay channels, rather than just $B^+ \rightarrow K^+ J/\psi(\rightarrow e^+e^-)$. Although the marginal gain in performance decreases with the number of samples—convergence to stable performance, the choice of 1/4 testing/training ratio has been made in section 3.2, because with higher testing/training ratios (tested for 3/7) the performance degraded. This effectively means that increasing the training sample size (from 70% to 80% of the full dataset) has improved the score, hence the stable performance point of neither of the algorithms has been reached. This implies that a further increase in training data could improve the performance [17].
From figure 5 it is clear that the portion of correctly clustered Bremsstrahlung photons is low compared to their true emitted amount. Some of them simply leave the detector, some of them do not comply with the thresholds set by the detector and some of them convert to e^- e^+ upstream of the magnet, and are subsequently bent away in the magnet. These are then not correctly clustered to their parent electron. Improving the clustering, and hence increasing the number of detected Bremsstrahlung photons would improve the reconstruction ability of the algorithms.
The SciKit documentation [18] claims that features of a dataset fitted by (polynomial) regression should not contain dependent variables: "When [...] the columns of the design matrix X [dataset] have an approximately linear dependence, the design matrix becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed target, producing a large variance." [18]. In this research project, mostly linearly independent features were used. However, the feature `Track_BremEnergy` is calculated using the other independent variables. One could argue that it may not be linearly dependent on the other features, however, the data pre-processing that introduces polynomial terms up to order r implies that the column in the dataset matrix corresponding to `Track_BremEnergy` should not be dependent on powers of features x_i up to order r, nor their cross-multiplication ($x_j \times x_k$...). Yet, the performance of the algorithm benefits from the inclusion of the `Track_BremEnergy` feature. An explanation for this surprising result may be given by examining the calculation of `Track_BremEnergy`.

4.4 Distribution irregularities

The ideal shape of the non-reconstructed residual sum of momenta $p_{recon} - p_{true}$ is expected to exponentially increase from $-\infty$ to 0, and then drop to 0, where it remains constant until $+\infty$. On the other hand, using the 0th order reconstruction $p_{recon} = p_{e^-} + p_{brem}$ should yield a more gradual decrease after $x = 0$, as now n may be more than the true value. Finally, the reconstruction using ML algorithms is expected to create a symmetric distribution around $x = 0$, with an exponential decrease on both ends. There are several discrepancies with the actual results in section 3.3. Firstly, the distribution of the non-reconstructed residual sum of momenta does drop to 0 on the positive side of the x-axis, however, there is an increase again at around $p_{recon} = p_{e^-} + p_{brem} = 3000$ MeV, which signifies that p_{recon} is over-estimating the true value. This can be seen in Figures 7a and 7b. Furthermore, both reconstruction algorithms have "bumps" that cause them to be asymmetrical in the y-axis. The peak of the polynomial regression is furthermore shifted to the right.
to explain (and correct) this unexpected behaviour, one can investigate the specific data points in the affected regions, and provide a physical explanation or in case the data points

correspond to non-relevant outliers, filter these points out.
Note that this behaviour does not occur when applying the filtering of $p_{brem} = 0$ events, as described in Section 3.4. It can therefore be concluded that at least partially responsible for the irregularities in the distributions in Section 3.3 are the events in which Bremsstrahlung is not accounted for.

5 Conclusion

The results of this research provide the following answer to the question "To what extent does the use of machine learning improve the electron momentum reconstruction performance of the LHCb?"

In general, both algorithms do improve the reconstruction capabilities of data points that lie on the extremities of the $p_{recon} - p_{true}$ distribution, which is indicated by the decrease in σ by 10% for the full dataset and 15% for the dataset with filtered 0 Bremsstrahlung events. Both ML algorithms also bring the average of the $p_{recon} - p_{true}$ distribution to zero, which implies that $p_{recon} = p_{true}$ on average. In other words, the observed under-estimation has been compensated by the ML algorithms.
On the other hand, we can see that the central part of the distribution is effectively spread out by both regression algorithms. In other words, for windows containing larger fractions of the data points, the Polynomial regression is favourable, whilst for small windows centred around x=0, the 0th order Bremsstrahlung addition offers a lower standard deviation.
The Polynomial Regression, however, offers a marginally lower standard deviation, whilst the DT has a lower MAE. There is a similar trade-off as with 0th order reconstruction and ML in the sense that Polynomial regression offers a lower standard deviation, hence is less spread out on the tails, but DT is more compact in the central region. Note that these differences are very marginal. Only when considering 99.8% of the data does the DT regression window $W_{99.8}(DT)$ surpass the size of the polynomial regression window $W_{99.8}(Poly)$. Due to the sensitivity of SD to the outliers, the 0.2 remaining 0.2% of the data on the very ends of the distribution is enough to contribute to the increased standard deviation of the DT over the polynomial regression.
It would therefore be beneficial to apply the ML algorithm only to data lying on the extremities of the distribution. The feasibility of this, however, is questionable as it is only with simulated data with known p_{true} where a plot of the residual $p_{recon} - p_{true}$, such as Figure 7a, can be made. If one could predict how far from p_{true} the value of p_{recon}, he could simply use that knowledge to "fix" the reconstruction in the first place.
Furthermore, since the computationally lengthiest part of supervised machine learning is the training, which only has to be done once, there should be no negative implications with employing supervised ML in the LHCb.
An improvement in e^- momentum reconstruction method over the current method leads to a better resolution, and lower uncertainty, which together with advancement in other areas of the LHCb detector may lead to the unambiguous proof of new physics behind the lepton universality violation.

References

[1] M. McNutt, "Self-Correction by Design," *Harvard Data Science Review*, vol. 2, dec 16 2020. https://hdsr.mitpress.mit.edu/pub/ivwnyg4w.

[2] J. Woithe, G. J. Wiener, and F. F. Van der Veken, "Let's have a coffee with the standard model of particle physics!," *Physics Education*, vol. 52, no. 3, p. 034001, 2017.

[3] S. Bifani, S. Descotes-Genon, A. R. Vidal, and M.-H. Schune, "Review of lepton universality tests in b decays," *Journal of Physics G: Nuclear and Particle Physics*, vol. 46, no. 2, p. 023001, 2018.

[4] M. van Veghel, *Pursuing forbidden beauty: Search for the lepton-flavour violating decays $B^0 \to e^{\pm}\mu^{\mp}$ and $B_s^0 \to e^{\pm}\mu^{\mp}$ and study of electron-reconstruction performance at LHCb*. PhD thesis, University of Groningen, 2020.

[5] M. Lubej, "Standard model." `http://www.physik.uzh.ch/groups/serra/StandardModel.html`, 2015. Accessed: 1-7-2022.

[6] N. Kruse, *Improving upstream electron identification with bremsstrahlung information*. PhD thesis, University of Groningen, 2020.

[7] K. Müller, L. Collaboration, *et al.*, "Tests of lepton flavour universality at lhcb," in *Journal of Physics: Conference Series*, vol. 1271, p. 012009, IOP Publishing, 2019.

[8] J. Lees, V. Poireau, V. Tisserand, J. G. Tico, E. Grauges, A. Palano, G. Eigen, B. Stugu, D. Brown, L. Kerth, *et al.*, "Measurement of branching fractions and rate asymmetries in the rare decays b→ k (*) + -," *Physical Review D*, vol. 86, no. 3, p. 032012, 2012.

[9] J.-T. Wei, P. Chang, I. Adachi, H. Aihara, V. Aulchenko, T. Aushev, A. Bakich, V. Balagura, E. Barberio, A. Bondar, *et al.*, "Measurement of the differential branching fraction and forward-backward asymmetry for b→ k (*) l+ l-," *Physical review letters*, vol. 103, no. 17, p. 171801, 2009.

[10] A. A. Alves Jr, L. Andrade Filho, A. Barbosa, I. Bediaga, G. Cernicchiaro, G. Guerrer, H. Lima Jr, A. Machado, J. Magnin, F. Marujo, *et al.*, "The lhcb detector at the lhc," *Journal of instrumentation*, vol. 3, no. 08, p. S08005, 2008.

[11] R. A. Nobrega, A. F. Barbosa, I. Bediaga, G. Cernicchiaro, E. C. De Oliveira, J. Magnin, L. M. de Andrade Filho, J. M. de Miranda, H. P. L. Junior, A. Reis, *et al.*, "Lhcb computing technical design report," 2005.

[12] A. Bates, A. Saavedra, C. Parkes, S. Viret, F. Marinho, and L. Dwyer, "Velo module characterisation: Results from the glasgow lhcb velo module burn-in," tech. rep., 2007.

[13] H. Bethe and W. Heitler, "On the stopping of fast particles and on the creation of positive electrons," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 146, no. 856, pp. 83–112, 1934.

[14] D. Berninghoff, J. Albrecht, and V. Gligorov, "Bremsstrahlung recovery of electrons using multivariate methods," tech. rep., Tech. Rep. LHCb-INT-2016-018. CERN-LHCb-INT-2016-018, CERN, Geneva, 2016.

[15] M. Clemencic, G. Corti, S. Easo, C. R. Jones, S. Miglioranzi, M. Pappagallo, and P. R. and, "The LHCb simulation application, gauss: Design, evolution and experience," *Journal of Physics: Conference Series*, vol. 331, p. 032023, dec 2011.

[16] "scikit-learn developers", "Scikit-learn linear models." `https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares`, 2022. Accessed: 1-7-2022.

[17] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, "Impact of dataset size on classification performance: an empirical evaluation in the medical domain," *Applied Sciences*, vol. 11, no. 2, p. 796, 2021.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.