Enhancing Primary Vertex Association with Machine Learning at the LHCb Experiment

Yukai Zhao
CERN; Heidelberg University (DE)

Supervised by
Abhijit Mathad
William Sutcliffe

Abstract

The correct association of the B meson is of great importance for accurate lifetime measurements. The PV-misassociation rate has been studied using Run 3 samples for semi-leptonic decays with varying numbers of neutrinos in the final state. Utilizing the association by selecting the minimal impact parameter in an event, a PV-misassociation rate of up to 5.5% was observed.

To improve the PV-misassociation rate, machine learning approaches have been explored. A simple Neural Network on PV-B binary classification resulted in an improvement of up to 1.3% in the decay channel the Neural Network is trained on and is able to generalize to the other decay channels of interest. As a second approach, a heterogeneous Graph Neural Network was implemented, as it more accurately represents the event structure. The heterogenous Graph Neural Network outperformed the simple Neural Network on the trained samples, however, it does not generalize to the other decay channels. Since the heterogeneous Graph Neural Network is still a work in progress, significant improvements are expected after further investigation.

Contents

Chapter 1

Introduction

The price study of B meson proves to be an interesting testing ground for the search for physics beyond the Standard Model. One critical aspect of this research is the analysis of $B^0_{(s)} - \bar{B}^0_{(s)}$ oscillations. For such analyses, the accurate determination of the primary vertex that produced the B is crucial for the correct decay time measurement. Furthermore for the tagging of the event using the opposite tagger, both B produced in an event need to be assigned to the same PV. One of the most prominent opposite taggers is the OS muon and OS electron tagger which uses the semi-leptonic decay mode of $B^0 \to D^{(*)-}\ell^+\nu$. Thus by looking at the charge of the leptons in the final state the flavour of the B can be determined.

The assignment of the primary vertex is performed by selecting the primary vertex which has the smallest impact parameter. However, this association is heavily influenced by the measured momentum of the B as well as the number of visible pp collisions in an event. Since the B^0 used for the oscillation measurement of flavour tagging is decaying semi-leptonic. The neutrinos in their final states are note reconstructed thus the momentum of the B^0 is deteriorated, leading to an increase in the impact parameter and thus misassociation

This issue was not significant during Run 1 and Run 2 of the LHCb experiment since on average only 1 visible pp collision is observed per bunch crossing. However, with the Upgrade 1 and starting from Run 3, the multiplicity has increased to an average of about 5 primary vertices per event. Consequently, the likelihood of misassociation using the assignment with the impact parameter rises significantly especially for the semi-leptonic decays. Looking ahead to Upfrase 2, an average of approximately 50 PVs is expected. Therefore, this report explores and summarizes alternative approaches to the minimum IP method.

Chapter 2

The LHCb experiment

The LHCb experiment, located at the Large Hadron Collider (LHC), is specifically designed to study bottom and charm decays to probe the SM of particle physics and search for beyond SM phenomena. The acronym LHCb stands for LHC *beauty*, referring to its focus on the beauty quark during the early stages of its existence. LHCb has been operating since 2010 during the data-taking periods Run 1 (2010-2012), Run 2 (2015-2018) and currently in Run 3 (2022-2025). With the start of Run 3, the instantaneous luminosity increased by a factor of 5 from $4 \times 10^{32}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$ to $2 \times 10^{33}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$. Looking ahead with the Upgrade 2 which takes place from 2035 to 2041, an increase of the instantaneous luminosity by another factor of 10 is expected.

2.1 The LHCb detector

The LHCb detector [1] is a single-arm forward spectrometer which covers a pseudorapidity η range of $2 < \eta < 5$. The pseudorapidity is defined as $\eta = -\log[\tan(\theta/2)]$ where θ is the polar angle with respect to the beam direction. The coordinate system used has the origin at the proton-proton interaction point, the z axis is aligned with the proton beam and pointing towards the muon system, the y axis pointing upwards and the x axis defining a right-handed system. Figure 2.1 shows the detector layout used for Run 3.

2.1.1 Track reconstruction

Trajectories of charged particles, called tracks, are reconstructed using the hits from the subdetectors. Different types of tracks are distinguished according to the subdetectors crossed, as shown in Figure 2.2:

- **Long tracks** require particles to cross the full tracking system and leave hits in all subdetectors.
- **Upstream tracks** are low momentum particles which leave hits in the VELO and UT but are swept out of the LHCb acceptance by the magnetic field.

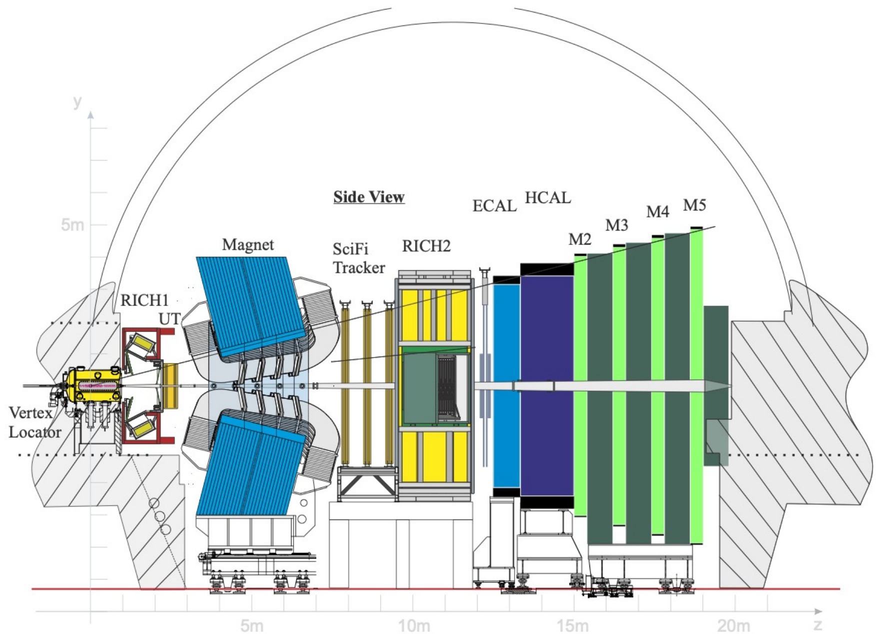


Figure 2.1: Layout of the LHCb detector for Run 3 [1].

- **Downstream tracks** are reconstructed using hits in the UT and SciFi Tracker.
- **Velo tracks** only have hits inside the VELO and are used to determine the position of the primary *pp* collisions, a process known as primary vertex finding.
- **T tracks** are only formed using hits from the SciFi Tracker.

2.2 Monte Carlo simulations

The study presented in this report is based on Monte Carlo (MC) [2] simulations. MC simulations involve the random generation of data which aim at replicating the behaviour of data obtained through experiments. To meet these requirements, two packages from the LHCb software are employed: Gauss [3] and Boole [4].
Gauss oversees the generation of the initial particles and the simulation of their transport through the LHCb detector. The simulation of the initial *pp* collision is carried out by the MC event generator Pythia [5]. The subsequent decays and time evolution of particles produced in the Pythia step are delegated to the EvtGen [6] framework. Afterwards, the Geant4 [7] toolkit is responsible for simulating the physical processes which the particles undergo as they traverse the detector.
The Boole program replicates the various responses of the subdetectors to hits and performs digitization, converting the data into the same format as that provided by the experimental electronics. The resulting files are referred to as MC simulations in Digi format. After digitization, real data and MC follow the same processes. The HLT reconstruction and

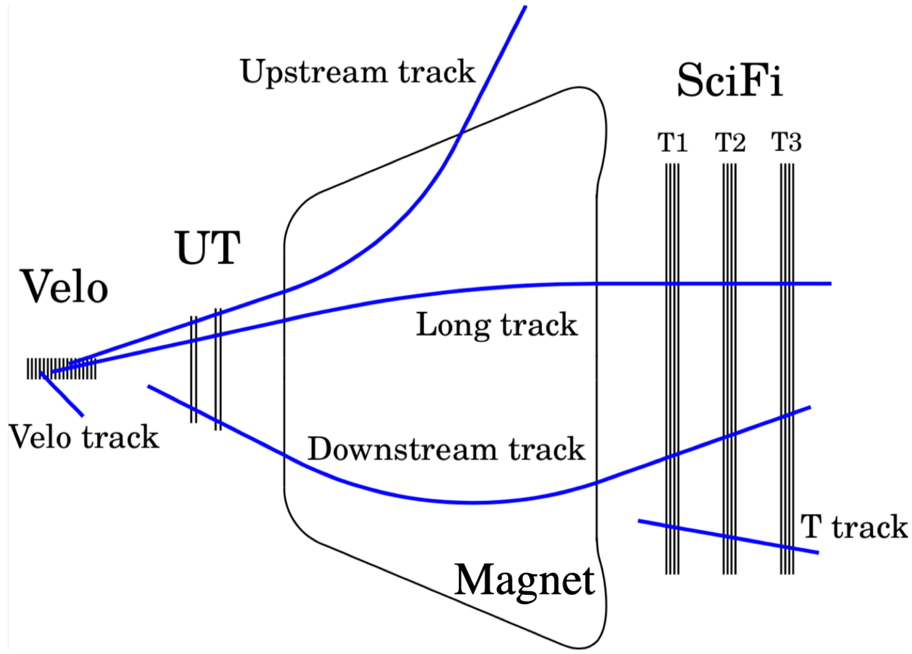


Figure 2.2: Track types in the LHCb detector bending plane. Adapted from [1].

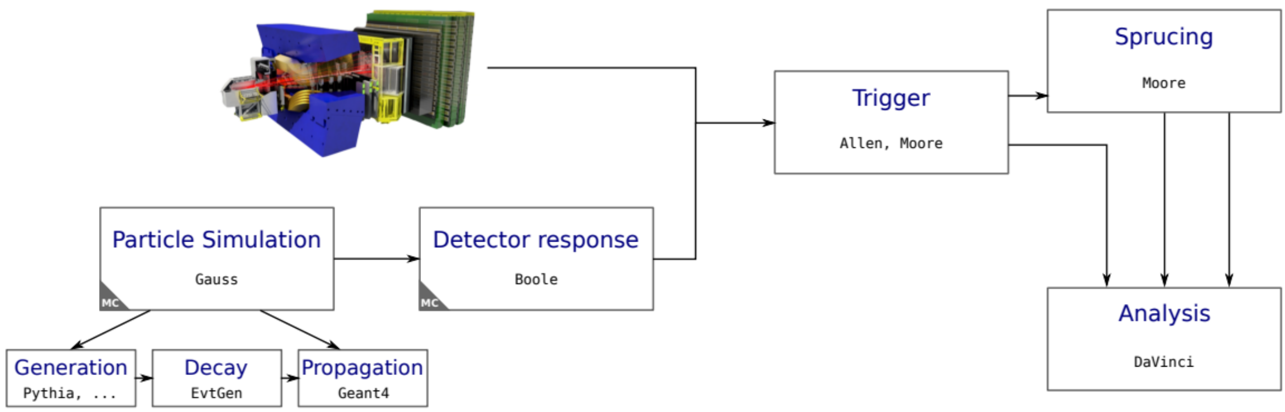


Figure 2.3: Schematic representation of the LHCb upgrade data flow and the related LHCb application, with an emphasis on simulation [2].

selection are carried out by the MOORE framework. Afterwards, the tupling is handled by the DAVINCI framework. The complete data flow for MC simulations is depicted in Figure 2.3.

Chapter 3

Introduction to machine learning

In high-energy physics, classical machine learning algorithms, such as (boosted) decision trees, have become an established analysis tool. In recent years, more advanced algorithms, such as Deep Neural Networks, Graph Neural Networks, and other architectures, have shown promising results in various research fields, making their application in high-energy physics increasingly interesting

3.1 Neural Network

A Neural Network (NN) can be seen as a fit function $f_\theta(x)$ with a large number of model parameters θ. Like in traditional fitting methods, these parameters are optimized by numerically minimizing a loss function $\mathcal{L}$, which measures how well the network's predictions match the true values. The choice of the loss function depends on the specific problem since NNs are applied to various tasks such as classification or regression. The process of minimizing the loss function is called network training and requires a training dataset $(x, f)_j$. The function $f_\theta(x)$ can be represented as a sequence of smaller modules, known as layers. In a NN, information is passed from layer to layer in a process called message passing, where the data is progressively processed and transformed. This process can be expressed as

$$x \to x^{(1)} \to x^{(2)} \to \cdots \to x^{(N)} \equiv f_\theta(x) \tag{3.1.1}$$

Hence, the general structure of a network consists of N layers, including an input layer $x \to x^{(1)}$, an output layer $x^{(N-1)} \to x^{(N)}$, and $N-2$ hidden layers. The message passing between two layers $x^{(n-1)}, x^{(n)}$ in a fully connected or dense network is given by

$$x^{(n-1)} \to x^{(n)} := W^{(n)} x^{(n-1)} + b^{(n)} \tag{3.1.2}$$

where W is a $D \times D$ matrix, referred to as the network weights, and b is a D-dimensional vector known as the bias. Both the weights and bias are combined in the model parameter vector θ and are trainable parameters. In general, the dimensions of neighbouring layers do not need to be the same.

Splitting the vector $x^{(n)}$ into its D entries, defines the nodes, which form our network layer

$$x_i^{(n)} = W_{ij}^{(n)} x_j^{(n-1)} + b_i^{(n)} \qquad i = 1, \ldots, D \tag{3.1.3}$$

Since the transformations between layers are affine transformations, meaning that a combination of affine layers still results in an affine transformation, non-linear transformations called activation functions are added between the layers to introduce complexity to the NN. The simplest activation function is the Rectified Linear Unit (ReLU), defined as:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{3.1.4}$$

or the sigmoid activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1.5}$$

Thus, the message passing between two layers is given by:

$$x^{(n-1)} \to x^{(n)} := \mathrm{ReLU}\left[W^{(n)} x^{(n-1)} + b^{(n)}\right] \tag{3.1.6}$$

By following the arrows in eq. (3.1.1) and using eq. (3.1.6), the forward pass of the NN is defined. Thus, by passing through the training data, the loss can be calculated and averaged over the training data. To find the best model parameters, an iterative procedure called backpropagation of the loss is employed. It computes the gradient of the loss function with respect to each weight in the network. These gradients are then used to update the weights via an optimization algorithm like gradient descent, to minimize the loss function. However, the backpropagation step on a large dataset can be computationally expensive, so it is typically performed over small, random subsets of the training data. This method is called stochastic gradient descent, and these subsets are known as minibatches. The network parameters are updated after each minibatch, following:

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \alpha \left\langle \frac{\partial \mathcal{L}^t}{\partial \theta_j} \right\rangle \tag{3.1.7}$$

where α is the learning rate. Training a NN is performed over multiple iterations of all minibatches of the training dataset, with each iteration called an epoch, continuing until convergence is reached.

Chapter 4

Classical primary vertex association

The correct assignment of a particle to its *pp*-collision point, known as primary vertex (PV), is crucial for many analyses, such as lifetime measurements for oscillation studies.
Since a significant fraction of B mesons decay into a final state that includes a lepton and its corresponding neutrino, following lepton number conservation, these decays are referred to as semi-leptonic decays. As the neutrinos are not detectable, their momenta are missing in the reconstruction of the B meson's momentum. Hence, the reconstructed momentum does not precisely point to the PV which produced the B but is instead slightly deviated. This deviation did not pose a significant issue during Run 1 and Run 2 of the LHCb experiment due to the low multiplicity of PVs in each bunch crossing, which was approximately 1. However, with the start of Run 3 after Upgrade 1, the number of visible *pp*-collisions has increased to approximately 5. Consequently, it is expected that the effect of miss association in semi-leptonic decays is no longer negligible, which will be investigated in this chapter. Furthermore, with Upgrade 2 and starting with Run 5, it is estimated that the number of visible *pp* collisions will increase by a factor of 10, hence, significantly increasing the misassociation rate.

4.1 Data sets

In this study, Monte Carlo (MC) simulations of five different decay channels are employed to study the primary vertex misassociation rate of the B meson. The channels were chosen to include a varying number of neutrinos in the final state to investigate their impact on the misassociation rate. Additionally, two reference channels $B^+ \to J/\psi K^+$, which are fully reconstructed, were selected. In the first fully reconstructed decay mode, the J/ψ decays into an electron pair. This channel is used to study the effect of the reduced electron momentum resolution, caused by Bremsstrahlung photon emissions, on the misassociation rate. In the second channel, the J/ψ decays dimuonic and serves as a cross-check, since the B momentum is well reconstructed, and thus, no significant miss association is expected. All decay channels and their number of events after truth matching are summarized in Table 4.1, with charge conjugation implied throughout the report.

Table 4.1: Decay channels of interest with $D^{*+} \to D^0\pi^+$ and $D^0 \to K^+\pi^-$.

Sample	Events	information
$B^0 \to D^{*-}\mu^+\nu_\mu$	32656	1 missing ν
$B^+ \to D^0\tau^+(\to \bar{\nu}_\tau\pi^+\pi^-\pi^+)\nu_\tau$	2198	2 missing ν
$B^0 \to D^{*-}\tau^+(\to \bar{\nu}_\tau\mu^+\bar{\nu}_\mu)\nu_\tau$	23677	3 missing ν
$B^+ \to J/\psi(\to e^+e^-)K^+$	284	electron resolution
$B^+ \to J/\psi(\to \mu^+\mu^-)K^+$	4770	fully reconstructed

4.2 PV association by minimizing impact parameter

Currently, the association of the primary vertex (PV) to the B meson is performed by selecting the PV with the smallest impact parameter IP among all PVs produced in an event. The impact parameter represents the shortest distance between the B's trajectory and the PV, as illustrated schematically in Figure 4.1. The impact parameter is calculated as follows:

$$\mathrm{IP} = \frac{\left|(\vec{v_0} - \vec{r}) \times \vec{p}\right|}{\left|\vec{p}\right|} \tag{4.2.1}$$

where $\vec{p}$ denotes the momentum of the B, $\vec{v_0}$ the position of the PV, and $\vec{r}$ the endvertex position of the B. The association is then performed by selecting the PV with the smallest impact parameter value. The selected PV is defined as the best primary vertex (BPV) and the method is referred to as the minIP method.

The determination of the impact parameter is heavily reliant on the momentum of the B meson, which leads to a larger impact parameter in semi-leptonic B decay channels, due to the missing neutrino momenta in the B momenta reconstruction. This effect is illustrated in Figure 4.2a, where the impact parameter distributions to the true PV are compared between the fully reconstructed decay channel $B^+ \to J/\psi(\to \mu^+\mu^-)K^+$ and the decay channel with one missing neutrino, $B^0 \to D^{*-}\mu^+\nu_\mu$. The impact parameter in the decay channel with the missing neutrino is significantly larger, as expected.

However, the increase in the impact parameter value in decay channels with a neutrino in

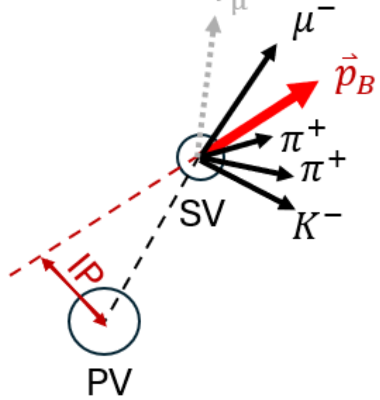
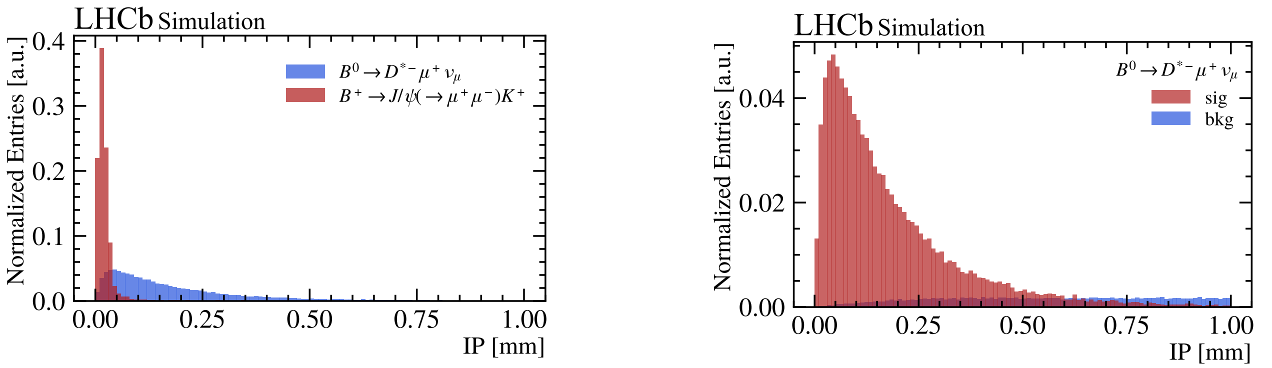


Figure 4.1: Schematic depictions of the decay topology in the x-z plane, along with the impact parameter.



(a) Impact parameter distribution to the correct PV for $B^0 \to D^{*-}\mu^+\nu_\mu$ and $B^+ \to J/\psi(\to \mu^+\mu^-)K^+$.

(b) Impact parameter distribution between PV which produced the B and background PV.

Figure 4.2: Impact parameter distribution of different channels and between signal PV and background PV.

the final state does not imply that the association of the PV using the minIP method is highly inefficient. As shown in Figure 4.2b, the impact parameter distribution for the signal PV (the PV that produced the B) and the background PV (the PV that did not produce the B) are compared. For the signal PV, there is a clear peak at low IP values, whereas the tail of the background PV distributions, which is peaking at an IP value of $\sim$ 12 mm, is only visible. This demonstrates that the impact parameter remains a powerful tool for discriminating between PVs when selecting the correct one. It is important to note that the selection is not performed on the absolute IP value of a PV but by the minimal value in the event.

4.2.1 Miss-assocation rate using minimal impact parameter

Utilizing the minIP method to associate B to the PV, the misassociation rate can be calculated. The misassociation rates for the different decay channels are summarized in Table 4.2. A misassociation rate of up to 5.5% was observed in the neutrino final state decays. Thus, indicating that the misassociation has a non-negligible impact. In the case of the two neutrino final state decay, a significantly lower misassociation rate is observed compared to the other neutrino final state decay channels. It is hypothesized that the high multiplicity of the three-prong tau decay leads to a small phase space for the neutrinos. Hence, the neutrinos produced in this decay are in general softer and, therefore, the momentum of the B meson is less deteriorated, leading to better reconstruction and, consequently, a more accurate PV association of the B. Furthermore, because of the higher multiplicity in tracks, the endvertex of the B should be better reconstructed. Additionally, it is suspected that the muons in the other semi-leptonic decays are quite soft, thus more momenta is carried away by the neutrinos. These reasons might lead to the smaller impact parameter distribution for the two neutrino final state decay as seen in Figure 4.3. However, this needs to be further studied in detail. For the fully reconstructed channels, where J/ψ decays into an electron pair or dimuonic, a clear improvement in the misassociation rate is observed, as expected, since there are no missing particles in the final state. The misassociation rate in the electron pair

Table 4.2: Misassociation rates for the different decay modes using the **minIP***method, where the uncertainties are statistical.*

Sample	Events	Misassociated	Misassociation rate [%]
$B^0 \to D^{*-}\mu^+\nu_\mu$	32656	1792	5.49 ± 0.04
$B^+ \to D^0\tau^+(\to \bar{\nu}_\tau\pi^+\pi^-\pi^+)\nu_\tau$	2198	60	2.73 ± 0.08
$B^0 \to D^{*-}\tau^+(\to \bar{\nu}_\tau\mu^+\bar{\nu}_\mu)\nu_\tau$	23677	1307	5.52 ± 0.05
$B^+ \to J/\psi(\to e^+e^-)K^+$	284	5	1.76 ± 0.15
$B^+ \to J/\psi(\to \mu^+\mu^-)K^+$	4770	14	0.29 ± 0.01

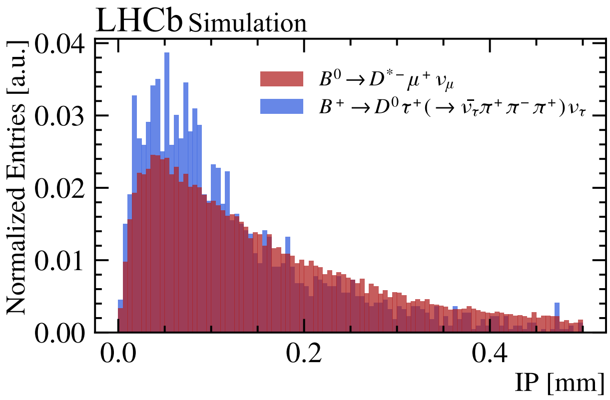


Figure 4.3: Impact parameter distribution of the one neutrino and the two neutrino final state. The two neutrino tend to have smaller impact parameters and thus a high association rate

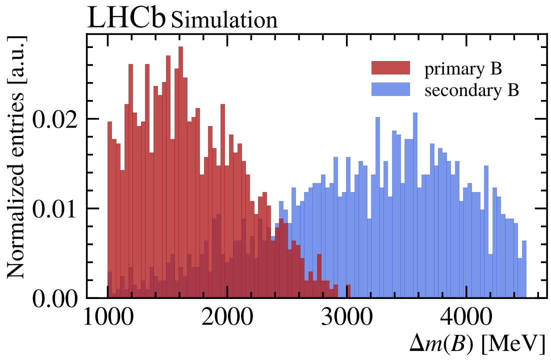
final state is slightly higher due to the worse momentum reconstruction of electrons which is mainly contributed to the emission of Bremsstrahlung photons.

4.3 Secondary B matching

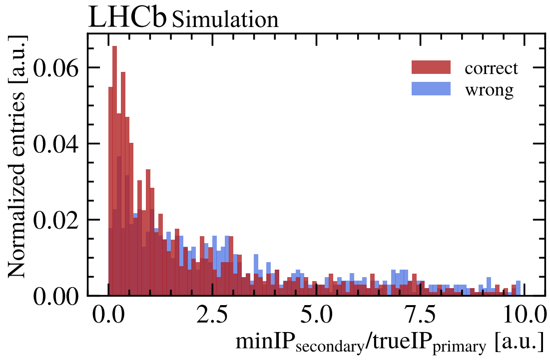
Since B mesons are always produced in pairs, both should ideally be reconstructed to the same PV. If one B decays semi-leptonically and the other decays fully reconstructable, the association of the semi-leptonic B decay could potentially be improved by looking at the secondary B. Thus by matching the B an improvement is expected.

As a proof of concept, no specific decay channels are required and all B mesons are considered. However, the secondary B must have at least three final state particles, which grand grandmother or less should be the B, due to limitations in the simulation. With these criteria, only ~ 6% of the events in the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay channel includes a secondary B meson in the event, corresponding to 2050 events out of 32656.

Since this could lead to a slight boost of PV association the secondary B mesons are evaluated how well they are reconstructed. This has been done by calculating the missing mass of both the primary B, which decays semi-leptonically and the secondary B, which is used for the matching. The resulting missing mass distributions of the primary and secondary B are shown in Figure 4.4a, where the secondary B is significantly worse reconstructed. Additionally, the ratio of the impact parameter of the best PV associated with the secondary B



(a) Missing mass distribution of the primary and secondary B.



(b) BPV impact parameter of the secondary B as a fraction of the true PV impact parameter of the primary B.

Figure 4.4: Primary and secondary B mission mass distributions (left). Fraction between impact parameter of secondary B BPV and true PV of the primary B, which gives a quantity to see if the secondary B is better reconstructed.

to the impact parameter value of the true PV of the primary B was estimated, as shown in Figure 4.4b. The distributions were separated into two categories: one where the secondary B BPV matches the true PV and one where it does not. No clear difference is visible between the distributions. Thus, the feasibility of this method is not possible and may even worsen the association accuracy.

Chapter 5

Machine learning approach

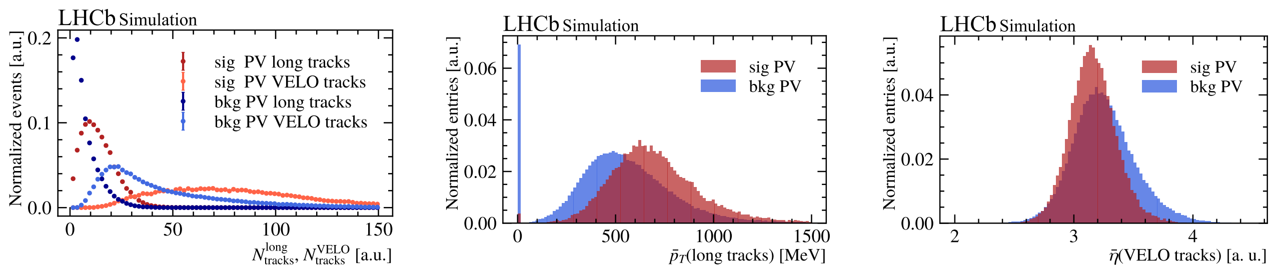
The production of b quarks differs significantly from that of light quarks. Light quarks are mainly produced through nonperturbative quantum chromodynamics (QCD) processes. However, due to the high mass of the b quark, this process is highly suppressed. Additionally, the valence quark distribution of the incoming beam particles contains no b quarks [8]. As a result, the $b\bar{b}$ production is dominated by hard parton-parton interaction in perturbative QCD. Consequently, a signal PV is expected to exhibit different features on the event level compared to a background PV. This difference suggests that an improvement of the misassociation rate is achievable by including event-level features.

Since the difference is expected at the pp collision, only prompt tracks should exhibit distinguishing power. This is evident in Figure 5.1a, which shows the distribution of the number of long and VELO tracks between signal and background PV. The signal PV produces significantly more long and VELO tracks. Furthermore, the mean transverse momentum $\bar{p}_T$ of the long tracks associated with the PV, and the mean pseudorapidity $\bar{\eta}$ distributions of the VELO tracks associated with the PV is shown in Figure 5.1b and Figure 5.1c, where it becomes evident that the kinematic variables of the track yield differentiating power. In the mean transverse momentum distribution, an additional peak at zero mean transverse momentum is observed, which arises from PVs reconstructed without long tracks, thus having no transverse momentum.

To improve the association rate of semi-leptonic B decays, a Neural Network (NN) approach is employed to incorporate additional event-level features, aiming to reduce the misassociation rate.

5.1 Neural Network

The general structure of a Neural Network (NN) can be found in Section 3.1. Since a NN always requires a fixed input dimension, this poses a challenge for our problem, as the number of primary vertices varies in each event, and even within a single PV, the number of tracks and VELO tracks can vary significantly. To address this, zero-padding would be applied so that all events are adjusted to have the maximum number of PVs, and each PV



(a) Distribution of the number of VELO tracks and long tracks for PVs which produced B and for PV which did not produce B.

(b) Mean transverse momentum distribution of the long track associated with the PV.

(c) Mean pseudorapidity distribution of the VELO tracks associated with the PV.

Figure 5.1: Different distribution comparison between PV which produced a B (red) and PV which did not produced a B (blue).

is padded to have the maximum number of tracks. However, this would lead to a network that is unnecessarily large and complex, making it difficult for the NN to effectively learn the association.

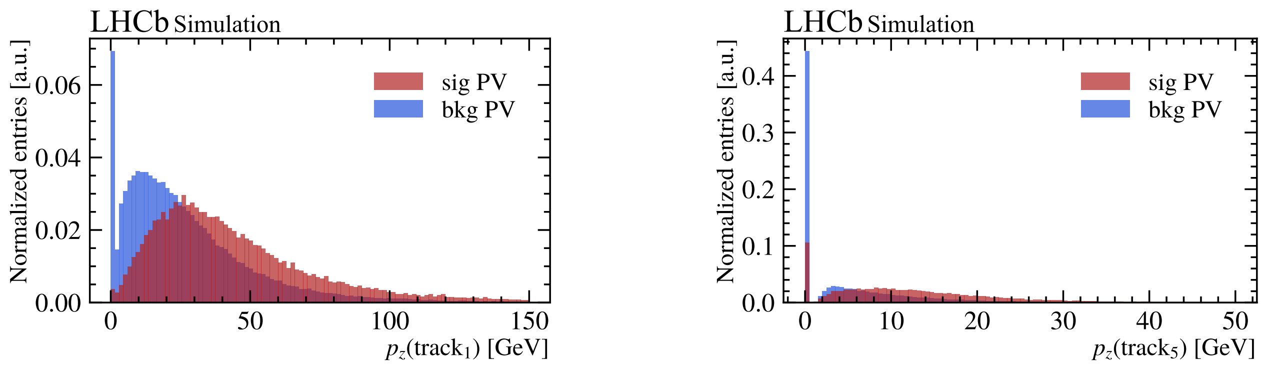
Therefore, a simpler approach has been chosen. The NN is designed to classify whether a given PV with its tracks matches the given B or not. Thus the problem is reduced to a binary classification of a PV being a signal or background PV. In the end, the association is performed, by combining the NN output of all PV-B combinations in an event and selecting the PV with the highest output value as the BPV.

5.1.1 Track selection

Since prompt tracks provide a significant discrimination power between signal and background PV, as shown in Figure 5.1, it is crucial to include their feature in the NN to boost the association rate. However, not all tracks can be included due to the constraints of the selected architecture, making it necessary to select the most relevant tracks.

One important factor to consider is ghost tracks, which are falsely reconstructed, hence non-physical tracks. Therefore, the ghost rate is determined for prompt tracks of different track types between signal and background PVs. The results, summarized in Table 5.1, show a clear significant increase in ghost rate between signal and background PVs across all track types. Consequently, a cut on the ghost probability of 0.07% is required to all tracks, significantly reducing the ghost rate. Since the association of tracks is performed using the minIP method, it is of interest to examine the misassociation rate of the prompt tracks as well. The impact parameter is calculated for long tracks using the reference point closest point to the beam pipe, while for upstream and downstream tracks the first measurement point is used. Comparing these results to the truth information, the misassociation rates are obtained and summarized in Table 5.2. Based on the high ghost rate and high misassociation rate, it becomes evident that downstream tracks should not be considered. Therefore, only long and upstream tracks are selected, given their high association and low ghost rates.

Another constraint is avoiding making the NN too complex and the fixed input feature



(a) Highest p_T long track. *(b) 5th highest p_T long track.*

Figure 5.2: z momentum distribution of the associated tracks of signal and background PV.

dimension. Consequently, only a limited number of tracks are used. The selection has been chosen only to consider the five highest p_T long tracks in a PV. This choice is based on several factors. First, as shown in Figure 5.1b, signal PVs tend to have higher p_T tracks on average. Second, in a PV more long tracks are produced, reducing the need for excessive zero-padding of non-existent tracks. Lastly, limiting the selection to only five tracks ensures that all selected tracks provide discrimination power, as demonstrated in Figure 5.2 while keeping the zero-padding at a moderate level (10% for the fifth track in signal PVs.).
It might be worth considering the inclusion of both long and upstream tracks simultaneously, however, this could complicate the selection process and would require further studies. Additionally, VELO tracks are not considered in this approach to simplify the NN used for this classification task.

5.1.2 Feature augmentation

Since the impact parameter is an efficient variable for selecting the right PV, it is expected to be an import an important feature for the NN as well. However, the impact parameter has intrinsic challenges which need to be challenged.
First, the impact parameter is highly correlated with the flight distance, which biases the lifetime measurements. Therefore, PV unbiasing was performed to tackle the bias introduced

Table 5.1: Ghost rate for different track types between signal and background PVs evaluated on a subset of the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay channel.

		no cut			$\mathcal{P}_{\text{ghost}} < 0.07$		
Track type	PV	Tracks	Ghosts	rate [%]	Tracks	Ghosts	rate [%]
Long	sig	9755	59	0.6	8561	9	0.1
	bkg	47748	8432	17.7	35775	1169	3.3
Upstream	sig	1581	2	0.1	1284	1	0.08
	bkg	6795	639	9.4	5213	127	2.4
Downstream	sig	432	43	10.0	367	18	4.9
	bkg	6347	4903	77.2	2532	1263	49.9

Table 5.2: Misassociation rate for different track type prompt tracks with ghost probability less than 0.07% *evaluated on a subset of the* $B^0 \to D^{*-}\mu^+\nu_\mu$ *decay channel.*

Track type	Tracks	Misassociated	Misassociation rate [%]
Long	43158	1128	2.61 ± 0.02
Upstream	6369	66	1.04 ± 0.02
Downstream	1618	1129	69.8 ± 3.6

by selecting the PV with minIP [9]. Hence, a more uncorrelated parameter to the flight distance is preferred.

Second, the NN tends to assign higher weights to features which significantly contribute to the classification task. Since the impact parameter values for signal PVs are typically much smaller than those for background PVs, the NN assigns large negative weights and large biases to the impact parameter, which are then penalized by the L2 regularization imposed by the weight decay. Therefore, a distribution where the signal values are larger than the background values is preferred, as it makes the learning process easier for the NN.

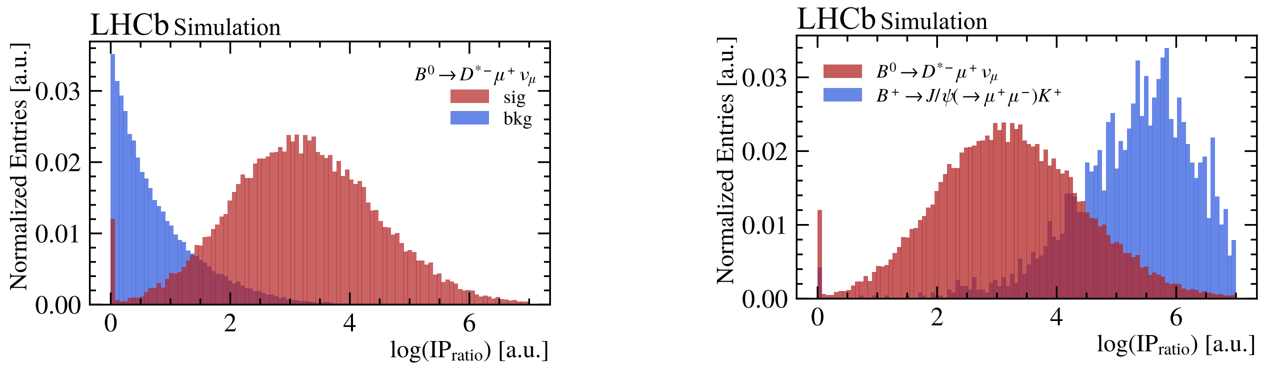
Additionally, for fully reconstructed channels, the impact parameter is even smaller compared to the decay channels with neutrinos in the final state, as seen in Figure 4.2a, leading to the requirement of even larger weights and biases to keep the same influence on the NN. Thus the NN might fail to generalize the PV selection to other decay channels.

As a result, the NN might struggle to learn a proper representation which effectively addresses all these issues. To mitigate these problems, an augmentation of the impact parameter is performed by utilizing the impact parameter ratio $\mathrm{IP}_{\mathrm{ratio}}$ instead of the impact parameter. The impact parameter ratio is defined as

$$\mathrm{IP}_{\mathrm{ratio}} = \frac{\sum \mathrm{IP}}{\mathrm{IP} \cdot \mathrm{nPVs}} \tag{5.1.1}$$

where the mean impact parameter value of the event is calculated and divided by the impact parameter of the PV of interest. This modification offers several advantages. First, it significantly reduces the correlation to the flight distance. Using the Pearson product-moment correlation coefficient, a correlation of 0.1 is obtained between $\mathrm{IP}_{\mathrm{ratio}}$ and the flight distance, compared to 0.5 for the impact parameter and flight distance. A value of 0 indicates no correlation, while a value of 1 indicates a high correlation. Second, the distribution now shows higher values for signals compared to the background. Additionally, the fully reconstructed channels have higher impact parameter ratio values than the semi-leptonic decay channels, making it easier for the NN to generalize. Lastly, by using the impact parameter ratio a relationship between the PV-B pairs within the event is added for the PV-B classification. This adjustment should make it easier for the NN to learn this feature effectively.

To prevent numerical instabilities in the NN and to achieve faster convergence, the features f_i are normalized to the order of $\mathcal{O}(1)$. The normalization is performed by subtracting the



(a) Impact parameter ratio distribution between PV which produced the B and background PV.

(b) Impact parameter ratio distribution to the correct PV for $B^0 \to D^{*-}\mu^+\nu_\mu$ and $B^+ \to J/\psi(\to \mu^+\mu^-)K^+$.

Figure 5.3: Impact parameter ratio distribution of different channels and between signal PV and background PV.

mean and dividing through the standard deviation

$$f_i = \frac{f_i - \langle f_i \rangle}{\sigma(f_i)} \tag{5.1.2}$$

However, for impact parameter related features, this process has been skipped to artificially enhance their importance.

5.1.3 Feature selection

A wide variety of features were tested using feature ablation, where each feature is individually set to zero to assess its impact on the model's accuracy. Furthermore, the flight distance and the endvertex position of the B are excluded despite showing discrimination power, due to the biasing of lifetime measurements. The chosen features are listed in Table 5.3, where $\sigma(z_{PV})$ is the uncertainty of the z position measurement of the PV, the number of tracks and mean distributions are independent of the constraints on the number of track features and $\vec{\text{dofm}}$ gives the distance between the primary vertex and the point of the first measurement. The minIP$_{1,2,3}$ gives the three smallest impact parameter values in an event, which is zero-padded in the case of less number of PVs in an event than 3. This feature gives additional event-level features to improve the classification. In total the number of features are 72.

5.1.4 NN layout and Hyperparameter optimization

Choosing the best hyperparameter plays an essential role in obtaining the optimal performance. Therefore, a hyperparameter optimizer based on an evolutionary algorithm is employed, to efficiently scan a wide range of hyperparameters.
In this evolutionary algorithm, an initial population of ten NNs with random hyperparameters are initialized. After the training, the top-performing models are selected as elites and passed on to the next generation. These elites exchange hyperparameters with each other

Table 5.3: List of features used for the NN training.

PV	B	track
χ^2	$\mathrm{IP}_{\mathrm{ratio}}$	$\vec{p}$
NDOF	χ^2_{red}	$\mathcal{P}_{\mathrm{ghost}}$
$\sigma(z_{PV})$		χ^2
# long tracks		NDOF
$\bar{p}_T$(long tracks)		$\vec{\mathrm{dofm}}$
# VELO tracks		
$\bar{\eta}$(VELO tracks)		
$\mathrm{minIP}_{1,2,3}$		

and have a small probability of mutation, where one hyperparameter receives a new random value. Additional models with randomly initialized hyperparameters are added to the elites until a population of ten models is reached again. This process repeats for ten generations and the hyperparameters of the best-performing model are chosen.
The optimization results in a four-layer NN with the following structure

- Input Layer : Input dimension 72, Output dimension 67
- Hidden Layer: Input dimension 67, Output dimension 67
- Hidden Layer: Input dimension 67, Output dimension 10
- Output Layer: Input dimension 10, Output dimension 1

ReLU activation function Equation (3.1.4) are applied between layers to introduce non-linearities. A batch size of 2048 was chosen, and the model was trained using the Adaptive Moment Estimation (ADAM) optimizer with an initial learning rate of 0.001. To prevent overfitting, regularization methods were employed, including dropout layers with a dropout rate of 34% between all layers. These dropout layers ensure that each node has a 34% chance of being deactivated during training, preventing the node's information from being further propagated. Additionally, on the optimizer, a weight decay of 0.009 was applied in the form of an L2-Regularization on the loss to penalize large weights. Furthermore, a learning rate scheduler with a step size of 10 and a gamma value of 0.5 was used, meaning that for every 10 epochs, the learning rate was multiplied by the gamma value. The loss is then estimated using a binary cross-entropy (BCE) loss with a sigmoid layer Equation (3.1.5), defined as

$$\mathrm{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{5.1.3}$$

with y the true label and $\hat{y}$ the predicted label. he sigmoid layer is used to enable the neural network to learn a probabilistic output more easily since the sigmoid function maps the output to a value between 0 and 1.
Since the idea was to use the differentiating power on the event level for the signal PV and background PV discrimination the training was only performed on the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay

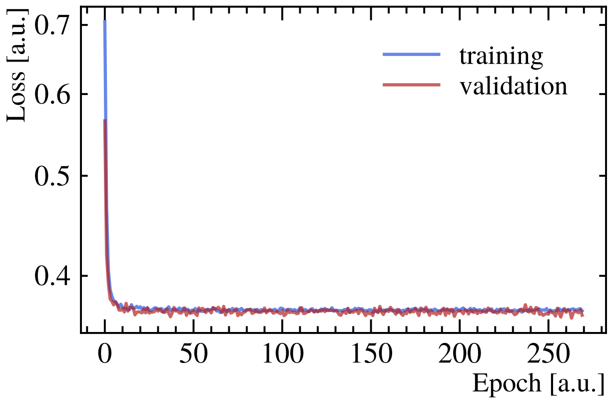


Figure 5.4: Loss curve, for both the dropout layers are active for comparable results.

channel to check for generalization and then applied to the other decay channels.
For the training process, the data set was split into three: 70% for the training, 20% for validation and 10% for testing. The testing set was used to check the misassociation rate of the events. The training was done over 270 epochs, reaching good convergence and no overfitting in the loss curve, as seen in Figure 5.4. The importance of all features can be seen in Appendix B, which are estimated by setting the feature of interest to zero. The most important feature is the $\mathrm{IP}_{\mathrm{ratio}}$, which remains throughout all the trained models. However, significant variation in the following features has led to the inclusion of additional features. While these may not have a significant impact on the model shown, they are of importantance in other models trained with the same settings.

5.1.5 NN results

As a first test of the performance of the trained NN on the $B^0 \to D^{*-}\mu^+\nu_\mu$ sample, classification accuracy between the signal and background PVs is evaluated. As shown in Figure 5.5a, the classifier output is presented, demonstrating the network is highly efficient in the selection between signal and background PVs. Using a simple threshold of an output value of 0.5 an estimation of the PV classification can be given, reaching an accuracy of 94.59% with the Receiver Operator Curve (ROC) of the NN in Figure 5.6.
Since the association using the PV-B classifier is performed by selecting the PV with the highest output value, the misassociation rate should be even lower. The distribution of the selected output correctly and wrongly associated events are shown in Figure 5.5b, inefficiency is indicated by a noticeable overlap between the correctly and wrongly associated distribution. Furthermore, it is noticeable that the misassociated events using the NN extend into the low-output region. Therefore, a threshold can be introduced to indicate that events below this region are not properly classifiable with the NN approach. The PV misassociation rate of both the minIP method and the NN as a PV-B classifier are summarized in Table 5.3. The uncertainty presented is of statistical nature. For the NN, the $B^0 \to D^{*-}\mu^+\nu_\mu$ sample exhibits a larger uncertainty due to the dataset being split into training, validation, and testing

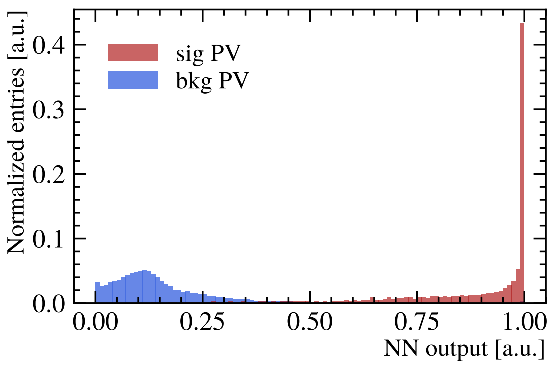
Table 5.4: PV-misassociation rate comparison between minIP *method and NN trained on* $B^0 \to D^{*-}\mu^+\nu_\mu$.

Sample	misassociation rate [%] minIP	NN
$B^0 \to D^{*-}\mu^+\nu_\mu$	5.49 ± 0.04	4.23 ± 0.13
$B^+ \to D^0\tau^+(\to \bar{\nu}_\tau\pi^+\pi^-\pi^+)\nu_\tau$	2.73 ± 0.08	2.44 ± 0.07
$B^0 \to D^{*-}\tau^+(\to \bar{\nu}_\tau\mu^+\bar{\nu}_\mu)\nu_\tau$	5.52 ± 0.05	4.73 ± 0.04
$B^+ \to J/\psi(\to e^+e^-)K^+$	1.76 ± 0.15	1.76 ± 0.15
$B^+ \to J/\psi(\to \mu^+\mu^-)K^+$	0.29 ± 0.01	0.29 ± 0.01

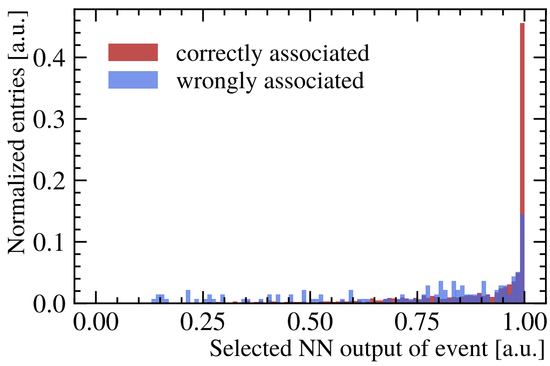
sets to obtain independent results, hence only 10% of the event are used.

The results clearly show a reduction in the misassociation rate across all semi-leptonic B decay channels, with a notable improvement of $\sim 1\%$ in the one and three neutrino final states. Even for the two neutrinos' final state, a slight improvement in the association is observed. The NN performs identically to the minIP method for the fully reconstructed decay channels. A possible explanation is that the $\mathrm{IP}_{\mathrm{ratio}}$ is larger for these channels, causing it to dominate over other features when using the same weights. As a result, the final selection is primarily driven by the $\mathrm{IP}_{\mathrm{ratio}}$, leading to the same performance between the NN and the minIP method. To address the problem, a model was trained on the $B^+ \to J/\psi(\to \mu^+\mu^-)K^+$ decay channel, resulting in an improvement in the $B^+ \to J/\psi(\to e^+e^-)K^+$ decay channel, with a misassociation rate of $1.41 \pm 0.12\%$. However, the PV association for the other decay channels worsened. As a potential solution, incorporating batch normalization between the layers might solve this issue, but further investigation is required.

Furthermore, it may be worth considering a preselection step, as some of the selected PVs result in negative flight distance in the z direction, which is unphysical and thus enhances the performance. Additionally, it might be worthwhile to apply a cut on the NN output which includes 99% of the signal PV and then apply it on the event level output. This should remove some Pv where the NN cannot perform a classification. These two approaches need to be further investigated.



(a) NN output of the PV-B classification.



(b) Selected NN for PV-B classification.

Figure 5.5: NN output values.

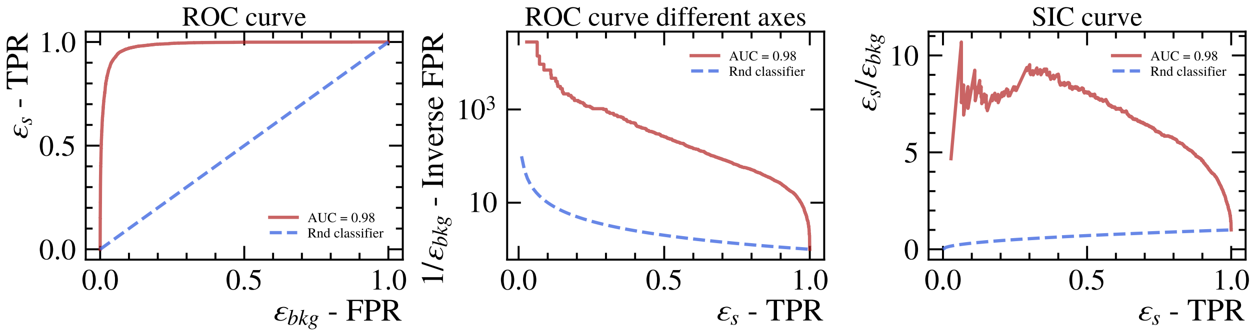


Figure 5.6: Receiver Operator curve of the trained NN.

5.2 Heterogeneous Graph Neural Networks

The primary issue with the current PV-B pair binary classification NN approach is the lack of connections between the events, despite adding features which try to catch it. Utilizing the maximum in the end to associate the B to the PV in the event is the only connection between all PVs. Additionally, the requirement of a fixed input dimension poses a significant challenge, since the number of tracks associated to a PV highly variate.
To address these issues an architecture is preferred which is able to take the whole event as an input and does not have issues with varying input dimensions.
A schematic representation of an event structure is shown in Figure 5.7. In an event, four different classes can be differentiated: PV, B, long tracks and VELO tracks, where each class contains its own features. Additionally, there are connections between the classes, which can hold features which are unique to the classes. This structure naturally motivated the application of a Graph Neural Network (GNN), where the classes are the nodes in the graph and the connection between the classes is known as the edges.
While typical GNNs assume that all nodes share the same feature set (homogeneous GNN), this assumption does not hold for this case, since PVs have different features compared to tracks for example. One could address this issue by utilizing zero padding and one hot encoding, which can be seen as a feature which is used to flag what kind of node it is. A more fitting approach would be to employ a heterogeneous GNN, which allows the definition of different node types and edge types, hence matching the graph-like structure of an event.

5.2.1 Intorduction to GNNs

The key concept of the GNN is message passing, which happens through a convolution process. Since message passing happens in similar manners for both homogeneous and heterogeneous GNNs, it will be explained in a homogeneous GNN setup. The extension to a heterogeneous GNN involves aggregating the message passing of the different edge types. In the message passing, a node aggregates information from its neighbouring nodes, which is given by the adjacency matrix A, and updates its own representations based on the

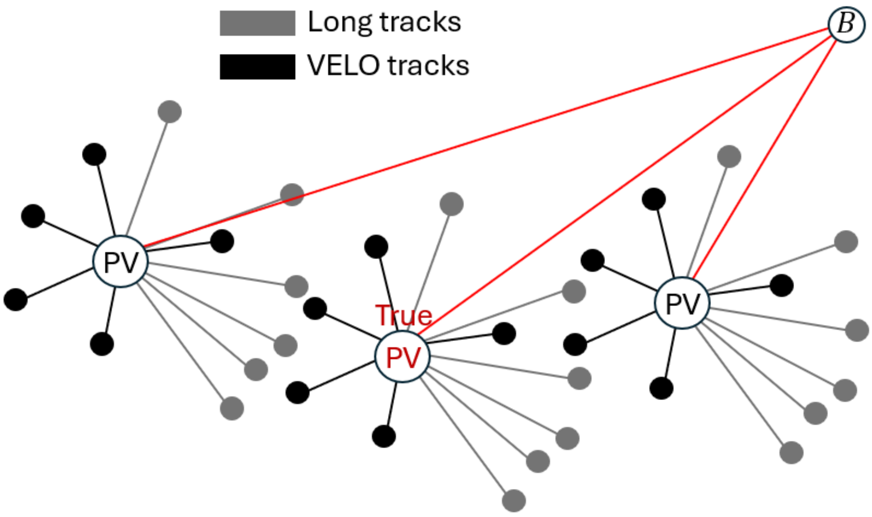


Figure 5.7: Schematic representation of an event emphasizing on the graph-like structure.

aggregated information. This process can be defined as :

$$x_i'^{(k)} = \sum_{\text{features l}} \sum_{\text{nodes j}} W_{ij}^{kl} A_{ij} x_j^l \tag{5.2.1}$$

where x is a node, W weights. The message passing can be written in the most general form, which is

$$x_i' = \text{ReLU } F_\theta[x_i, \square_{j \in N} \phi_\theta(x_i, x_j, e_{ij})] \tag{5.2.2}$$

where F_θ and ϕ_θ are trainable functions, $x_{i,j}$ node features and edge features e_{ij} with $\square$ an aggregation function such as mean or sum which combines the central node with a neighbourhood N. For each edge type, a separate convolution is necessary.
Three different convolutional layer types are used for this study.

Graph convolutional layer (GCNConv)

The graph convolutional layer [10] is the simplest layer which takes all connected nodes and performs the message passing. The message passing can be written as

$$x' = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} x W^{(l)}\right) \tag{5.2.3}$$

where $\tilde{A} = A + \mathrm{I}_N$ gives the adjacency matrix with an additional identity matrix I_N which indicated self loops, with N representing the neighbourhood, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ and $W^{(l)}$ are trainable weights, σ an activation function such as the ReLU.

Sampling and aggregation convolutional layer (SAGEConv)

In the sampling and aggregation convolution layer [11], not all nodes are used but a small subset is samples. This has the advantage of reducing the computational expansiveness significantly in large graphs. The convolution can be written as

$$x' = W_1 x_i + W_2 \cdot \text{mean}_{j \in N(i)} x_j \tag{5.2.4}$$

Graph attention convolutional layer (GATConv)

The graph attention convolutional layers [12] use the attention mechanism to select the nodes of interest. These attention mechanism then assigns them a weight which is used for the message passing. The message passing is then defined as:

$$x' = \sum_{j \in N(i) \cup \{i\}} \alpha_{i,j} W_i x_j \tag{5.2.5}$$

with α being the weight assigned by the attention mechanism. Furthermore, the number of heads can be adjusted, which determines how many sets of weights are assigned to a node by the attention mechanism. These heads can be thought of as analogous to feature maps.

Edge classification

The output after each convolution layer remains a graph since no pooling is applied. To associate the PVs using the heterogeneous GNN an edge classification is performed between the B and PV pairs. For this edge classification, the node features of the B and PV pair with their edge features are passed to a single-layer Multi-Layer Perceptron, which performs the binary classification.

Global features

The architecture of Graph Neural Networks allows the inclusion of global features, which can be seen as the total number of PVs in an event or the total number of tracks. Utilizing these global features an event can be described how "busy" they are. However, due to implementation issues global features are not included.

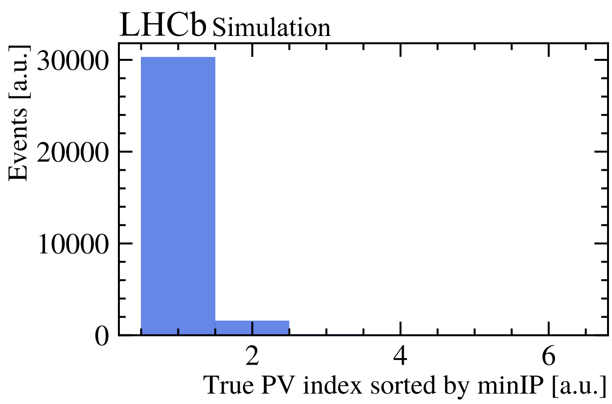


Figure 5.8: Index distribution of the true PV in minIP.

Table 5.5: Node features used for the heterogeneous GNN.

PV	B	VELO tracks	long tracks
$\vec{r}_{\text{pos}}$	$\vec{p}$	$\vec{r}_{\text{FM}}$	$\vec{p}$
χ^2	$\vec{r}_{\text{end}}$	η	χ^2
NDOF	χ^2		NDOF
# long tracks			$\mathcal{P}_{\text{ghost}}$
$\bar{p}_T$(long tracks)			$\vec{r}_{Fm}$
# VELO tracks			
$\bar{\eta}$(VELO tracks)			

Table 5.6: Edge features used for the heterogeneous GNN.

PV-PV	B-PV	VELO track-PV	long track-PV
$\Delta\vec{r}$	$\vec{FD}$	$\vec{\text{dofm}}$	$\vec{\text{dofm}}$
	IP	IP	IP

5.2.2 Feature selection

Reconstructing the graphs for an entire event can result in very large graphs. Hence, a preselection is necessary to simplify this problem. As the majority of the true PVs appear in the first of second index of the minIP distribution within an event, as shown in Figure 5.8, only two PVs are included in the graph. The PVs are selected by the smallest impact parameter value with the B, leading to a coverage of 99.5% of true PVs.

Since the study of the heterogeneous GNN is still ongoing, no constraints are currently imposed on the selected features for the GNN. The node features are summarized in Table 5.7, where $\vec{r}$ represents the position vector, and FM denotes the point of first measurement. The edge features are summarized in Table 5.6, with $\Delta\vec{r}$ representing the distance between two PVs.

5.2.3 Heterogeneous GNN setup

The heterogeneous GNN consists of three convolutional layers with a ReLU activation function in between two layers. Additionally, the graph has been made undirected. Since each edge type requires its own convolution, a total of seven convolutions are necessary per convolutional layer.

In the first layer, a GATConv with two heads is applied to all edges to select the important nodes through weights. The second and third layers are identical, where a GCNConv is used for the PV-PV edges as their number is quite small due to the preselection. For the edges PV to VELO and long tracks a SAGEConv layer is employed instead of a GATConv to reduce the complexity. For the remaining edges, a GATCONV with one head is applied.

The training is done with a batch size of 32 and a learning rate of 0.001 for the ADAM optimizer. The loss is determined using a binary cross-entropy loss with a sigmoid activation layer and can be seen Figure 5.9.

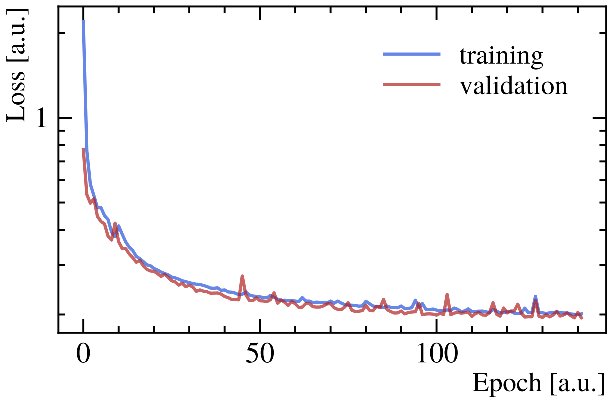


Figure 5.9: Loss curve of the trained GNN.

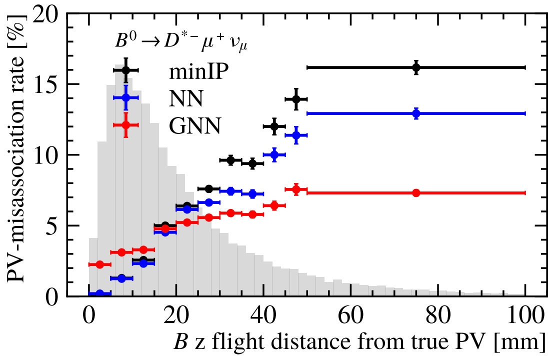
Table 5.7: PV-misassociation rate comparison between minIP *method and NN and heterogeneous GNN trained on $B^0 \to D^{*-}\mu^+\nu_\mu$.*

Sample	misassociation rate [%] minIP	NN	heterogeneous GNN
$B^0 \to D^{*-}\mu^+\nu_\mu$	5.49 ± 0.04	4.23 ± 0.13	4.12 ± 0.10
$B^+ \to D^0\tau^+(\to \bar{\nu}_\tau\pi^+\pi^-\pi^+)\nu_\tau$	2.73 ± 0.08	2.44 ± 0.07	4.00 ± 0.13
$B^0 \to D^{*-}\tau^+(\to \bar{\nu}_\tau\mu^+\bar{\nu}_\mu)\nu_\tau$	5.52 ± 0.05	4.73 ± 0.04	4.93 ± 0.05
$B^+ \to J/\psi(\to e^+e^-)K^+$	1.76 ± 0.15	1.76 ± 0.15	3.63 ± 0.31
$B^+ \to J/\psi(\to \mu^+\mu^-)K^+$	0.29 ± 0.01	0.29 ± 0.01	4.04 ± 0.08

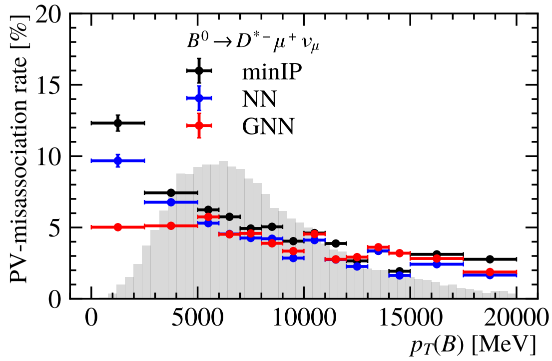
5.3 Results

The results of all three method are summarized in table 5.7. All machine learning algorithms are trained on the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay channel and applied on the other channels. Thus only a fraction of the data could be used for determineating the misassociation rate to keep it independent from the training. The heterogenous GNN results performs better compared to the simple NN approach. Sicne the heterogeneous GNN is still in a preliminary state it is expcetected that the final result will perform even better. However, a generalization is not achievable for the heterogeneous GNN.
Furthermore, the misassociation rate is shown as a function for different quanties in Figure 5.10. Multiple interesting quantities are observed. In the flight distance distribution the misassociation rate increases proportionally with the flight distance. HOwever utilizing the machine learning approach a significant reduction fo the missassociation rate for high flight distance values are obtained .THis is really of importance sine selecting the PV using minIP introduces a bias which needs to be undone using PV unbiasing. Thus this decreases in misassociaton rate in high flight distance regions is of great of importance. However a crossing of the heterogeneous GNN result is seen at an flight distance value of around 20, which is suspected to be traced back to the issue of the impact parameter value as discussed in Section 5.1.3. On the other side a more constant distribution of missassociaon rate would lead that the PV unbiasing process is not more necessary. However this needs to be further

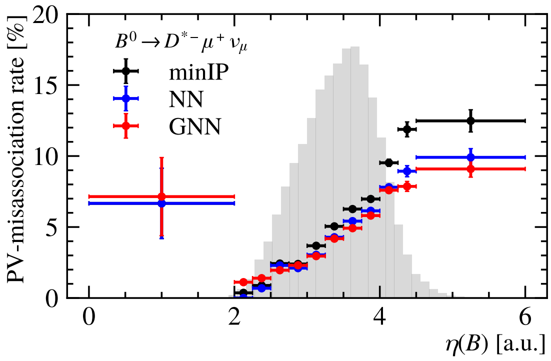
investigated. Furthermore with regard to the Upgrade 2 the distribution as a function of number of PVs in an event is shown, where it becomes evident that the machine learning approaches perform significantly better ath higher number of PVs in an event which is promising result wit regard to the application in RUn 5 and on. The distributions for the other channel are shown in Appendix C.



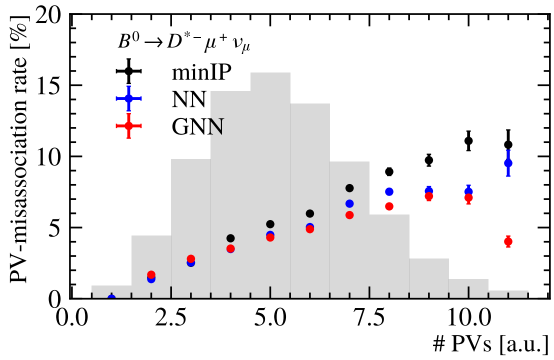
(a) PV-misassoiation rate as a function of the flight distance of the B.



(b) PV-misassoiation rate as a function of the transverse momentum of the B.



(c) PV-misassoiation rate as a function of the pseudo-rapidity of the B.



(d) PV-misassoiation rate as a function of the numbers of PVs in an event.

Figure 5.10: PV-misassociation rate as a function of different quantities for the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay channel.

Chapter 6

Conclusion

Through Monte Carlo simulation, the primary vertex misassociation rate has been determined for semi-leptonic and fully reconstructable B decay modes using the current algorithm of PV association which is based on selecting the minimal impact parameter value. For the fully reconstructable dimuonic decay mode, a misassociation rate of 0.29 ± 0.01 was obtained, as it was expected. In the case of the electron pair final state in the fully reconstructable decay mode, a misassociation rate of 1.76 ± 0.15 was obtained, which is explained by the worse electron momentum reconstruction due to the emission of bremsstrahlung photons. For the semi-leptonic case, a misassociation rate of $\sim 5\%$ is obtained for the one and three neutrino final states. In the case of the two neutrinos final state, a misassociation rate of 2.73 ± 0.08. The lower misassociation rate is hypothesized to be due to topological differences in the decay compared to the other semi-leptonic decays.

To improve the misassociation rate, a classical approach was first investigated by attempting to identify the secondary B which is produced in the $b\bar{b}$ production and afterwards matching the primary B to the same vertices. This approach would require that the secondary B is better reconstructed than the primary B. Although this approach should theoretically work, but limitations in the Monte Carlo simulations caused issues. Further investigation into this method is required.

Since $b\bar{b}$ production differs significantly from the production of light quarks, the PV that produced a B should differ from a PV that did not produce a B. Therefore machine learning approaches are investigated to capture this difference.
As an initial approach, a simple Neural Network was developed as a binary classifier to determine whether a given PV produced the B or not. This is performed for all PV in one event and then the one with the highest output is selected. The Neural Network has been trained on the $B^0 \to D^{*-}\mu^+\nu_\mu$ and applied on the other channels. Using this approach, a $\sim 1\%$ improvement in the PV-misassociation rate in the semi-leptonic decay channels was obtained while maintaining the same performance in the fully reconstructed channels. This approach of the Neural Network has the limitations that the full event information is not included

in the classification as well as the fixed input size, to utilize all tracks produced in a PV. This motivated the exploration of Graph Neural Networks, especially heterogeneous Graph Neural Networks. This approach has the advantage that different nodes can be defined, representing the PV, B, VELO tracks and long tracks. Furthermore, different edges can be defined between the node types. Thus, making the heterogeneous approach a suitable model.

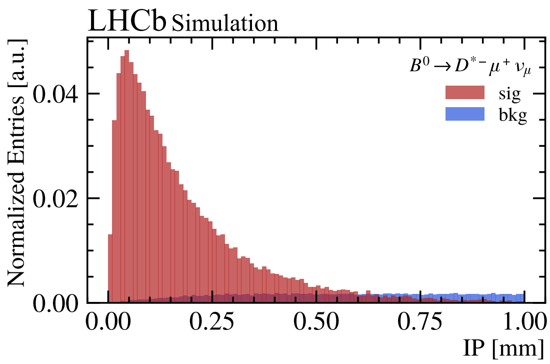
As the GNN is still a work in progress, a simplified setup was used, which only includes two PVs with the smallest impact parameter value in the event, which covers 99.5% of cases. Similar to the simple Neural Network approach it has been trained on the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay mode and applied on the other decay channels. Using this approach the PV-misassociation rate in the trained decay channel improved slightly further. However, a generalization to the other decay channels was not achieved, as their misassociation rate increased significantly. It is suspected that the impact parameter feature between B and PV causes this issue, but further investigation is necessary.

While a $\sim 1\%$ increase in the PV-misassociation rate may not seem significant, the machine learning algorithms perform significantly better at higher flight distance, with an improvement up to $\sim 7\%$ in the $B^0 \to D^{*-}\mu^+\nu_\mu$ decay channel. This improvement is of great interest for lifetime measurements. Furthermore, the machine learning approaches also perform significantly better in events with a higher number of PVs, which is of great importance considering Upgrade 2, where the number of visible Pvs per bunch crossing increases by a factor of ten.

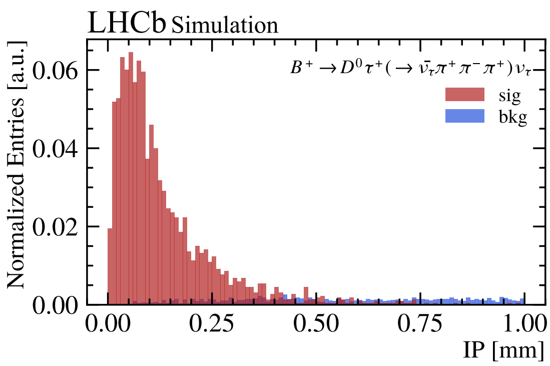
It has been clearly shown that machine learning approaches are of great interest in tackling the problem of PV-misassociation rate and with further study an improvement in the PV-misassociation rate using the heterogeneous Graph Neural Network is expected.

Appendix A

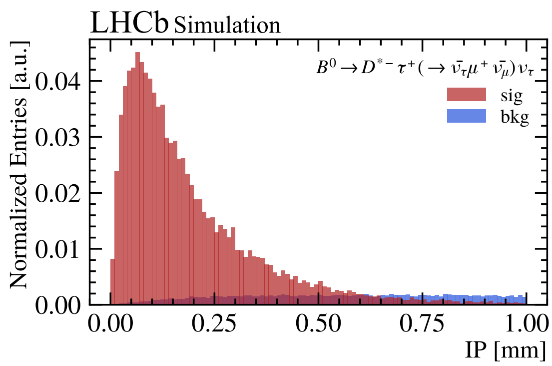
IP distribution



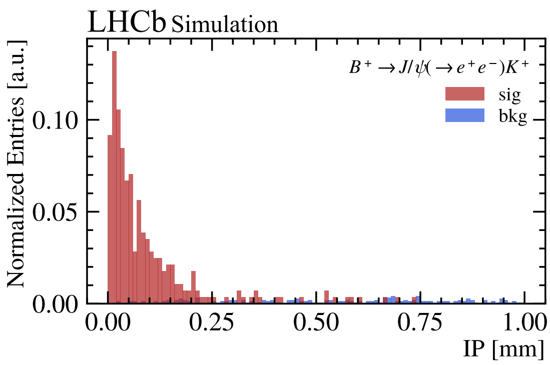
(a) Caption for Center Figure



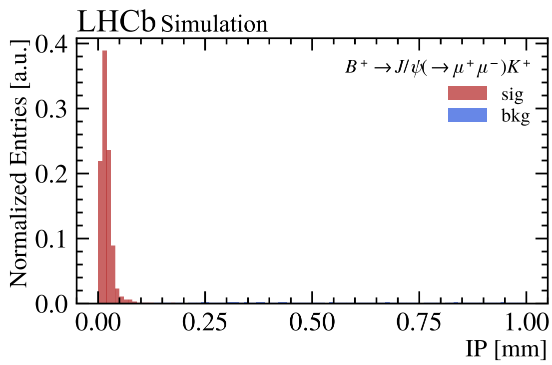
(b) Caption for Figure 1



(c) Caption for Figure 2

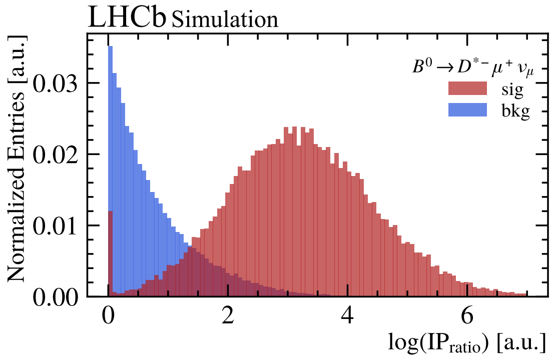


(d) Caption for Figure 3

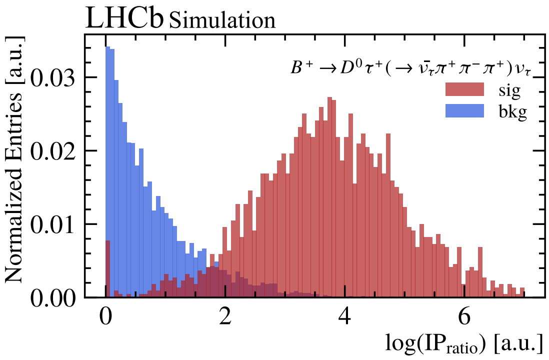


(e) Caption for Figure 4

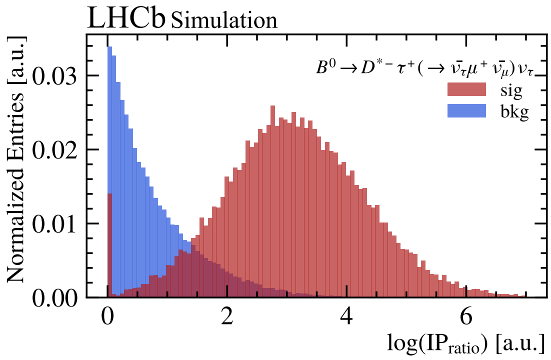
Figure A.1: Overall caption for the figure grid with a top center figure



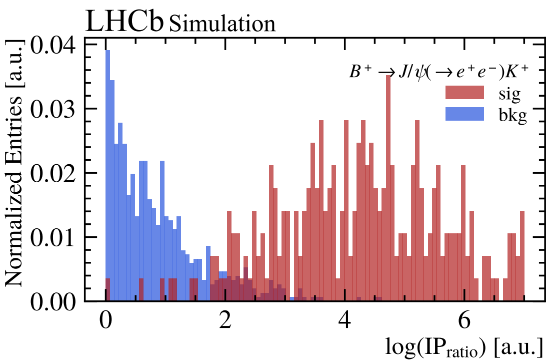
(a) Caption for Center Figure



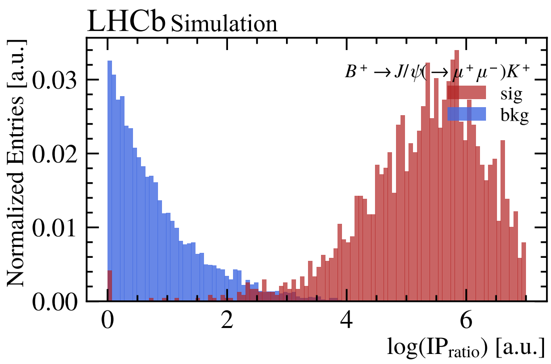
(b) Caption for Figure 1



(c) Caption for Figure 2



(d) Caption for Figure 3



(e) Caption for Figure 4

Figure A.2: Overall caption for the figure grid with a top center figure

Appendix B

Feature importance in NN

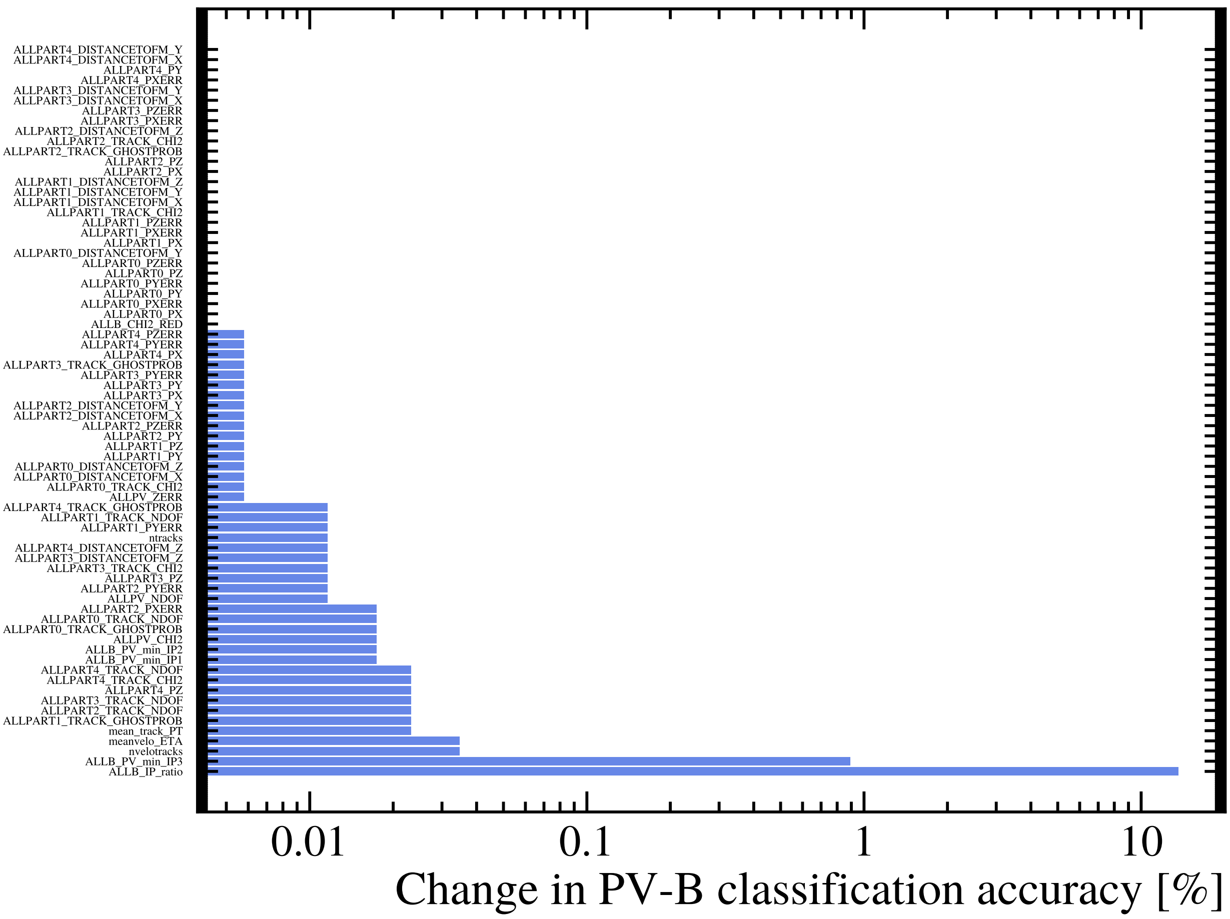
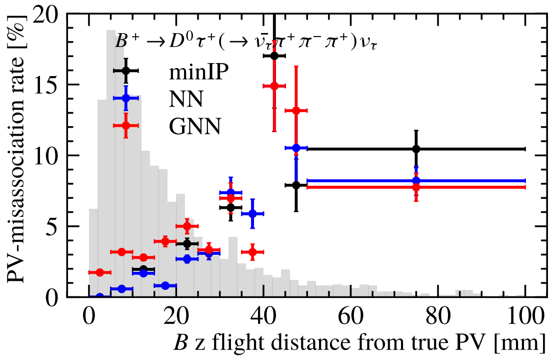


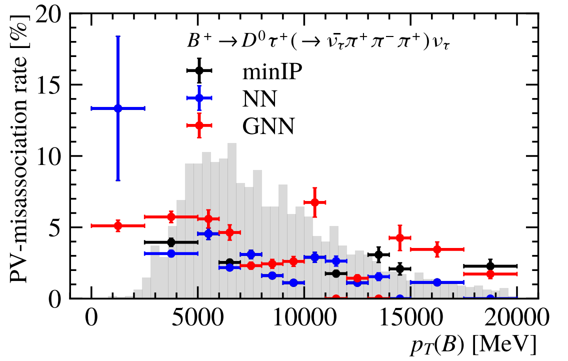
Figure B.1: Importance of the features used in the NN. The importance of features varies greatly for different models trained on the same dataset.

Appendix C

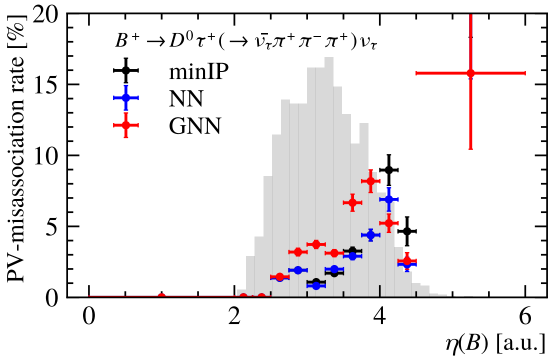
PV-misassociation rate distributions for different features



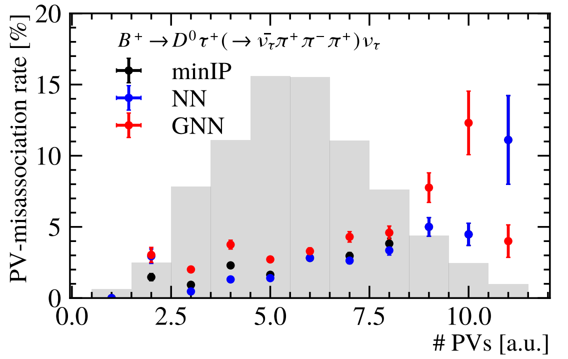
(a) PV-misassoiation rate as a function of the flight distance of the B.



(b) PV-misassoiation rate as a function of the transverse momentum of the B.

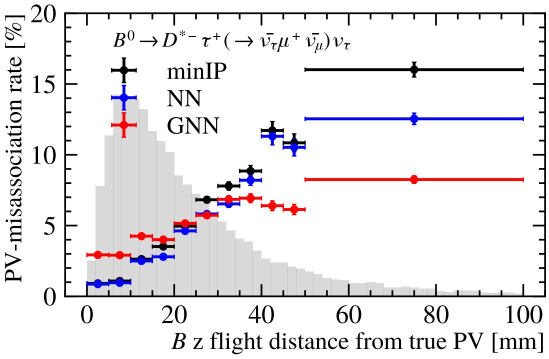


(c) PV-misassoiation rate as a function of the pseudo-rapidity of the B.

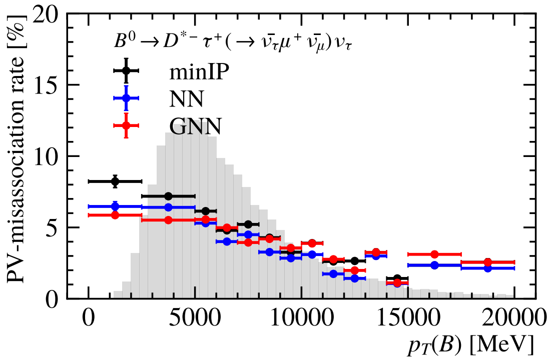


(d) PV-misassoiation rate as a function of the numbers of PVs in an event.

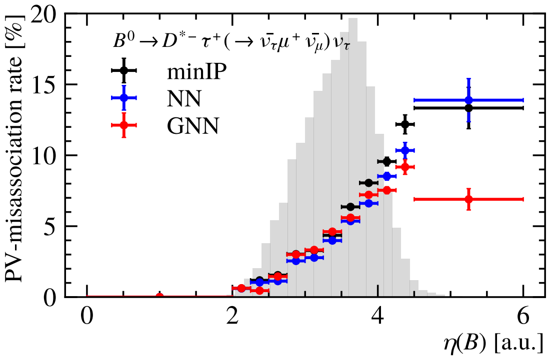
Figure C.1: PV-misassociation rate as a function of different quantities for the $B^+ \to D^0\tau^+(\to \bar{\nu}_\tau\pi^+\pi^-\pi^+)\nu_\tau$ decay channel.



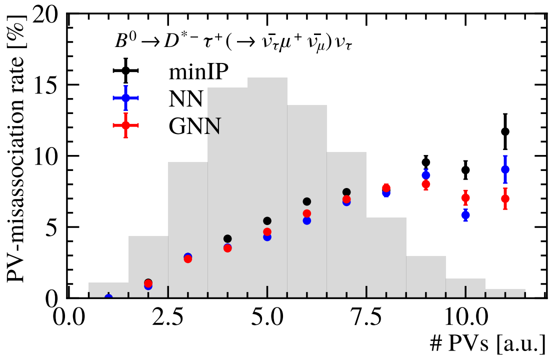
(a) PV-misassoiation rate as a function of the flight distance of the B.



(b) PV-misassoiation rate as a function of the transverse momentum of the B.

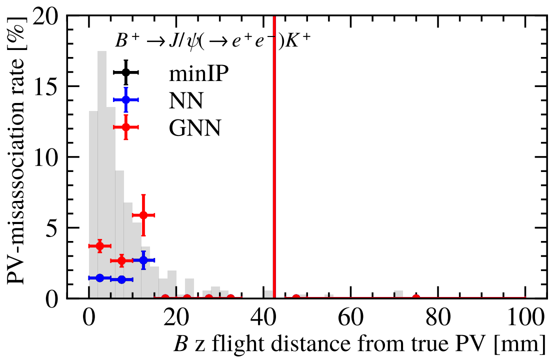


(c) PV-misassoiation rate as a function of the pseudo-rapidity of the B.

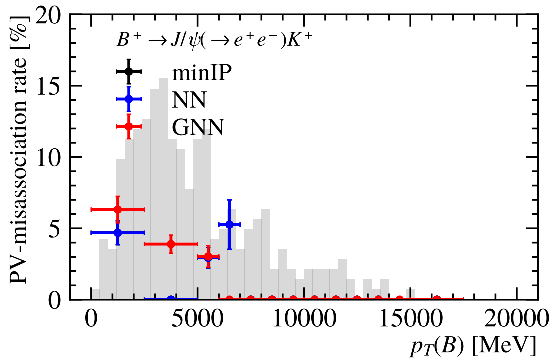


(d) PV-misassoiation rate as a function of the numbers of PVs in an event.

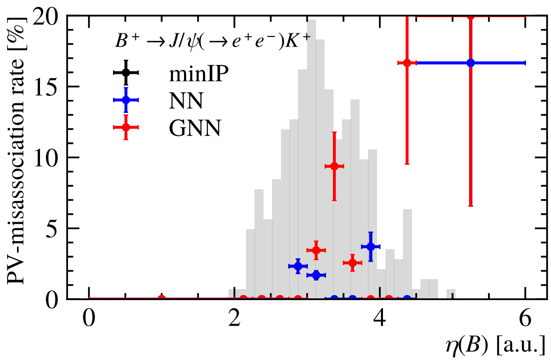
Figure C.2: PV-misassociation rate as a function of different quantities for the $B^0 \to D^{*-}\tau^+(\to \bar{\nu}_\tau\mu^+\bar{\nu}_\mu)\nu_\tau$ decay channel.



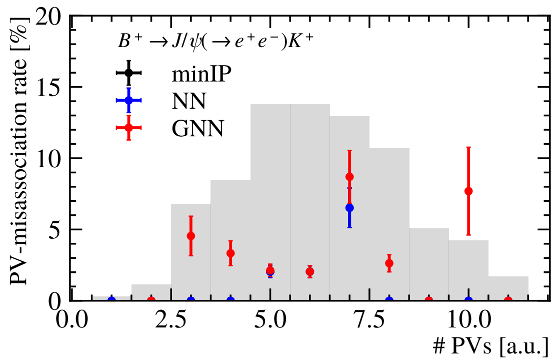
(a) PV-misassoiation rate as a function of the flight distance of the B.



(b) PV-misassoiation rate as a function of the transverse momentum of the B.

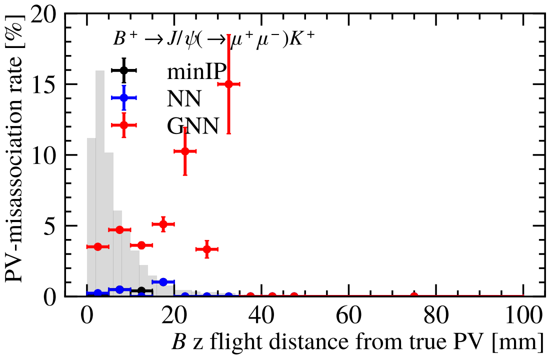


(c) PV-misassoiation rate as a function of the pseudo-rapidity of the B.

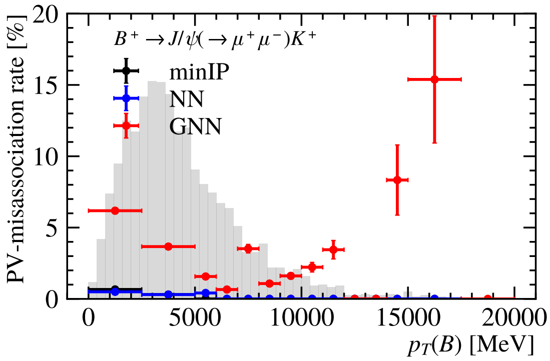


(d) PV-misassoiation rate as a function of the numbers of PVs in an event.

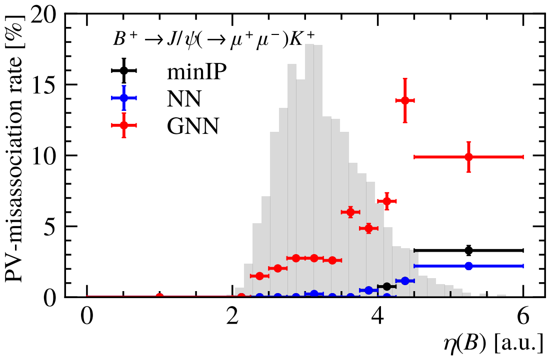
Figure C.3: PV-misassociation rate as a function of different quantities for the $B^+ \to J/\psi(\to e^+e^-)K^+$ decay channel.



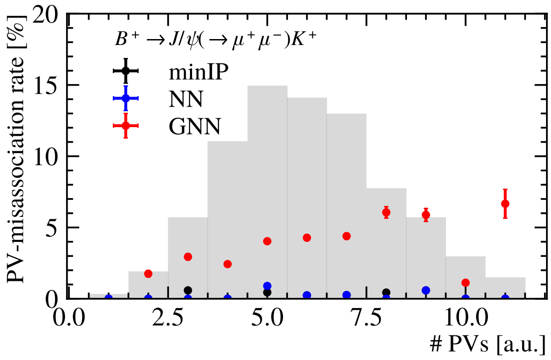
(a) PV-misassoiation rate as a function of the flight distance of the B.



(b) PV-misassoiation rate as a function of the transverse momentum of the B.



(c) PV-misassoiation rate as a function of the pseudo-rapidity of the B.



(d) PV-misassoiation rate as a function of the numbers of PVs in an event.

Figure C.4: PV-misassociation rate as a function of different quantities for the $B^+ \to J/\psi(\to \mu^+\mu^-)K^+$ decay channel.

Bibliography

[1] LHCb collaboration et al. *The LHCb upgrade I*. 2023. arXiv: 2305.10515 [hep-ex].

[2] C. Abellán Beteta et al. *Calibration and performance of the LHCb calorimeters in Run 1 and 2 at the LHC*. 2020. arXiv: 2008.11556 [physics.ins-det].

[3] M Clemencic et al. "The LHCb Simulation Application, Gauss: Design, Evolution and Experience". In: *Journal of Physics: Conference Series* 331.3 (Dec. 2011), p. 032023. DOI: 10.1088/1742-6596/331/3/032023.

[4] R. Antunes Nobrega et al. "LHCb Computing Technical Design Report". In: *Technical Design Report LHCb* (2005).

[5] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), pp. 159–177. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.01.024.

[6] David J. Lange. "The EvtGen particle decay simulation package". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 462.1 (2001). BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, pp. 152–155. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)00089-4.

[7] S. Agostinelli et al. "Geant4—a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8.

[8] E. Norrbin and T. Sjöstrand. "Production and hadronization of heavy quarks". In: *European Physical Journal C* 17 (2000), pp. 137–161. DOI: 10.1007/s100520000460.

[9] Matthew William Kenzie and Vladimir Gligorov. *Lifetime unbiased beauty and charm triggers at LHCb*. Tech. rep. Geneva: CERN, 2016.

[10] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG].

[11] William L. Hamilton, Rex Ying, and Jure Leskovec. *Inductive Representation Learning on Large Graphs*. 2018. arXiv: 1706.02216 [cs.SI].

[12] Petar Veličković et al. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML].