LHCb-PUB/CONF/INT/PAPER-2016-YYY
August 26, 2016

Fast neural-net based fake track rejection

Paul and others (Stephen Farry, Sascha Stahl, Michel DeCian, who else?)

Abstract

A neural-network based algorithm to identify fake tracks in the LHCb pattern recognition allows to use the same track reconstruction online (i.e. the second stage of the software trigger) and offline. A low fake rate reduces the load on the particle identification and the combinatorics of decay reconstructions. At the same time, the algorithm fits into the CPU time budget of the farms.

1 Introduction

The LHCb detector consists of subsystems designed to perform high efficiency tracking ($> 95\,\%$) with an excellent momentum resolution ($0.5\,\%$ for $p < 20\,\mathrm{GeV}/c$). Two Ring Imaging Cherenkov detectors provide precise particle identification. In Run II of the LHC, a new scheme for the LHCb software trigger allows splitting the triggering of the event in two stages, giving room to perform the alignment and calibration in real time. In the novel detector alignment and calibration strategy for Run II, data collected at the start of the fill are processed in a few minutes and used to update the alignment, while the calibration constants are evaluated for each run. This allows identical constants to be used in the online and offline reconstruction. The larger timing budget, available in the trigger, and a reduced computing resource demand of the event reconstruction result in the convergence of the online and offline track reconstruction. The same performance of the track reconstruction and PID are achieved online and offline. This offers the opportunity to optimise the event selection in the trigger with stronger constraints and including the hadronic PID. It additionally increases selection efficiencies and purity and reduces systematic uncertainties.

A key ingredient to achieve this ambitious goal is the reduction of the fake track rate prior to the particle identification and combinatorics of reconstructing particle decays in the second software trigger stage. A neural network, described in this note, is deployed to identify fake tracks, called the "ghost probability".

Terminology

To avoid ambiguity, the bare term "performance" is avoided. Instead, when referring to how well good tracks are separated from fake tracks, the term "physics performance" is used since it is the figure of merit on which physics analyses depend. The term "CPU performance" is used for the amount of computing resources needed to execute the algorithm proposed in this note. As benchmark for the latter, the cycle count of callgrind [1] is used. Effects of instruction caching and data caching are assumed small, approximately confirmed by wall clock time measurements. The cycles spent in other algorithms which are only called to compute input quantities to the ghost probability are accounted to the ghost probability, most notably this comprises the interpolation of tracks through active detector material to determine which channels should have a hit from the track – algorithms like the track fit, which would be executed anyways, are not accounted to the ghost probability.

The term "ghost probability" is used for both, the entire algorithm computing whether a track is considered a fake track or a real track, including the neural network, and for the numeric response of that algorithm. A selection requirement at the nominal working point of 0.4 is implied[1] when the ghost probability is referred to as a selection requirement.

[1] this corresponds to a fake track retention of $40\,\%$

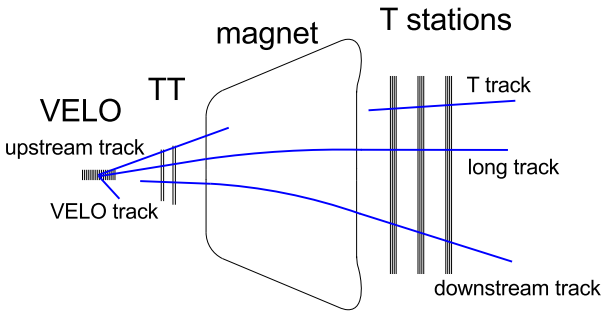

Figure 1: Illustration of the tracking system of LHCb, starting from the VELO around the collision point on the left, particles pass the TT, are deflected in the magnetic field of the dipole magnet and then detected in the T station detector (IT and OT). Different track types are reconstructed by different track finding algorithms. [2]

1.1 Track reconstruction

Owing to the design of the LHCb detector, which consists of tracking detectors mainly outside the magnetic field, charged particle tracks are in approximation straight line segments in the upstream part (VELO and TT) and in the downstream part (T stations). Figure 1 shows an overview of the different track types defined in the LHCb reconstruction: VELO tracks, which have hits in the VELO; upstream tracks, which have hits in the two upstream trackers; T tracks, which have hits in the T stations; downstream tracks, which have hits in TT and the T stations; and long tracks, which have hits in the VELO and the T stations. The latter tracks can additionally have hits in TT.

If a particle is reconstructed more than once, as different track types, only the track best suited for analysis purposes is kept. Hereby, long tracks are preferred over any other track type, upstream tracks are preferred over VELO tracks, and downstream tracks are preferred over T tracks.

Most analyses use long tracks because they provide the best momentum and spatial resolution among all track types. Unless otherwise stated, track reconstruction at LHCb refers to the reconstruction of long tracks. In a typical signal triggered event, around 60 long tracks are reconstructed. Other track types, such as downstream tracks, are used for the reconstruction of decay products of long-lived particles such as K^0_S mesons, or for internal alignment of the tracking detectors.

Tracks are fit with a Kalman filter fit. In addition to a global fit χ^2, separate contributions to the χ^2 from the downstream detectors (IT and OT), χ^2_D, and from the upstream detectors (VELO and TT), χ^2_U are computed. A large number of fake tracks results from wrong combinations of well reconstructed track segments in the upstream and downstream regions. These usually have good χ^2_D and χ^2_U but the additional contribution from matching the two segments, $\chi^2_M = \chi^2 - \chi^2_D - \chi^2_U$, is large for these "matching" fakes.

The Kalman fit has an outlier removal to account for individual detector hits which are not due to the reconstructed particle track. Beyond that, a special treatment for Outer

Tracker hits is in place. The readout electronics is designed to select only a single hit in each channel per bunch crossing; if two charged particles pass the same straw, a drift time measurement will only provided for one of them. To describe tracks in high occupancy OT modules, the drift time measurement can be ignored and only the information that a track went somewhere through the straw is used. This is decided for each straw–track combination individually if the hit residual is too large, similar to a standard outlier removal. This drift time suppression ensures that the track fit χ^2 is not biased to larger values for tracks in high multiplicity events, for tracks in the OT with respect to tracks in the IT, or for tracks in high occupancy modules close to the beam axis.

1.2 Previous works

An earlier version of the work presented here was already used in analyses of Run I data. The neural network was evaluated in the offline reconstruction to distinguish fake tracks from real particles' tracks [3] (used e.g. in [4]). The network was trained on all reconstructed tracks in simulated events with at least one $b\bar{b}$ pair produced in the pp collision.

The 22 input variables to the old ghost probability are the track fit χ^2, and the individual contributions $\chi^2_D, \chi^2_U, \chi^2_M$ and the corresponding degrees of freedom; the numbers of hits on the track in each tracking detector; the reconstructed track p_{T} and pseudorapidity; the track is interpolated through the detector to count in how many active strips/straws no hit is observed although the track passes through it ("expected hits"); and finally the occupancies of all tracking detectors.

There are separate networks for each track type, where input variables are removed if they are not defined for that track type (e.g. VELO hits for downstream tracks).

1.3 Training sample

The LHCb track reconstruction needs to be able to handle a wide range of LHC running conditions. At the time of preparing for data taking in 2015 it was not clear whether the LHC would operate at 25 ns bunch spacing or 50 ns bunch spacing. Simulations to prepare the track reconstruction were prepared for these scenarios:

- 25 ns bunch spacing, $\nu = 1.6$
- 25 ns bunch spacing, $\nu = 1.9$
- 50 ns bunch spacing, $\nu = 1.6$
- 50 ns bunch spacing, $\nu = 2.7$

The scenarios differ, for what concerns the track reconstruction, significantly in detector occupancy and spillover in the Outer Tracker. That may lead to different behaviours of fake track reconstruction and require different network trainings for different running conditions. The necessity for having different network trainings is assessed by training

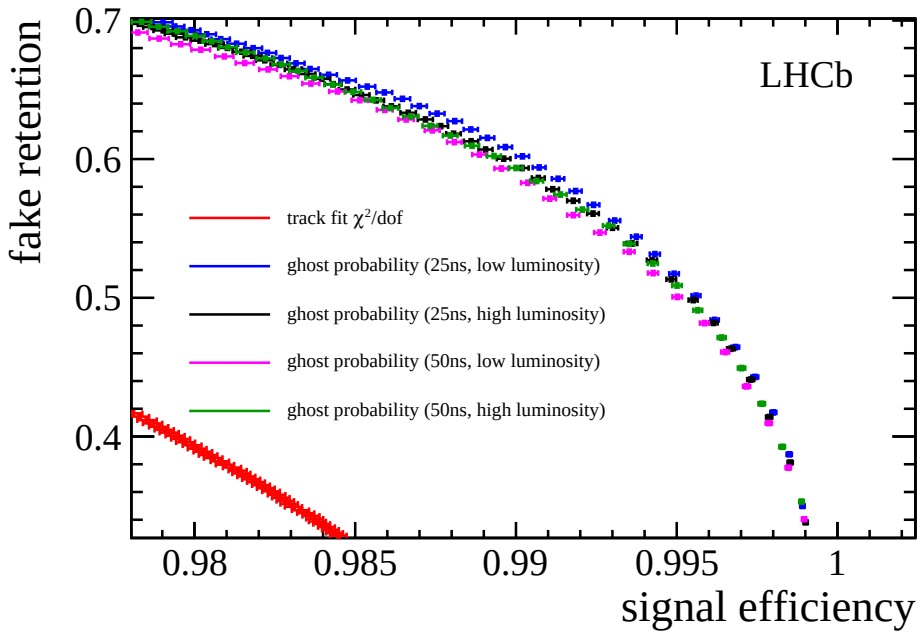

Figure 2: ROC curves for fake track discriminating neural networks, trained for different LHC running conditions, evaluated for 25 ns bunch spacing, $\nu = 1.6$. The error bars are statistical uncertainties only.

networks for each of the running conditions, with all other training parameters fixed, and evaluating the networks on one of the samples and their discriminating powers are compared. Figure 2 shows the ROC curves for the 25 ns sample at low pile-up. The discrimination powers of the four networks do not largely differ and thus for simplicity only a single network (trained at the favoured scenario of 25 ns bunch spacing at low pile-up) is deployed.

1.4 Network architecture tuning

As framework for the neural network, the TMVA package [5] is chosen since it is

- equipped with a root file interface for the training, which is the common data file format in LHCb software,
- commonly known in LHCb (ensuring future maintainability),
- and provides code generation for the trained network such that the network can be integrated into any C++ code without creating dependency on external libraries.

At the time of development, the $\tanh(x)$ function was a commonly used activation function in TMVA, while known as a computationally expensive function to be optimised for the LHCb pattern recognition [6]. Yet it is not the only possible sigmoid function [7] and consequently a custom activation functions have been added to TMVA [8].

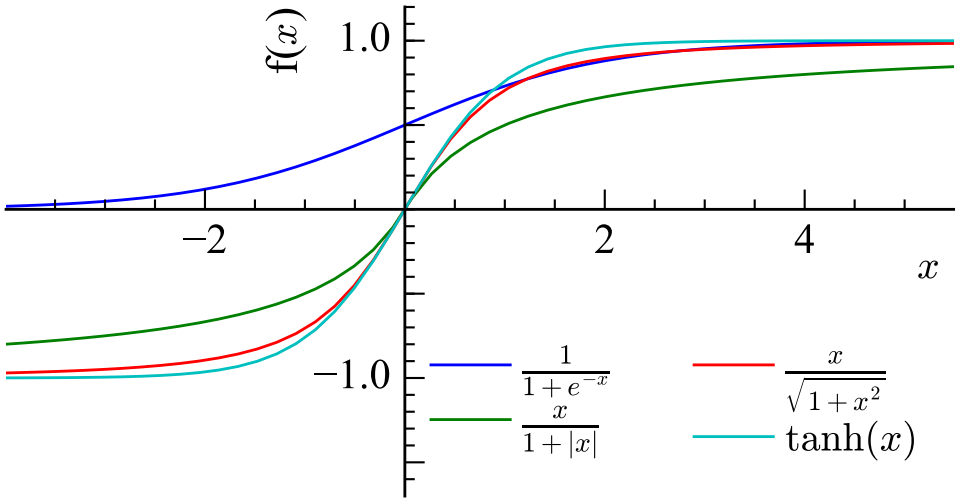

Figure 3: Functional shape of sigmoid functions.

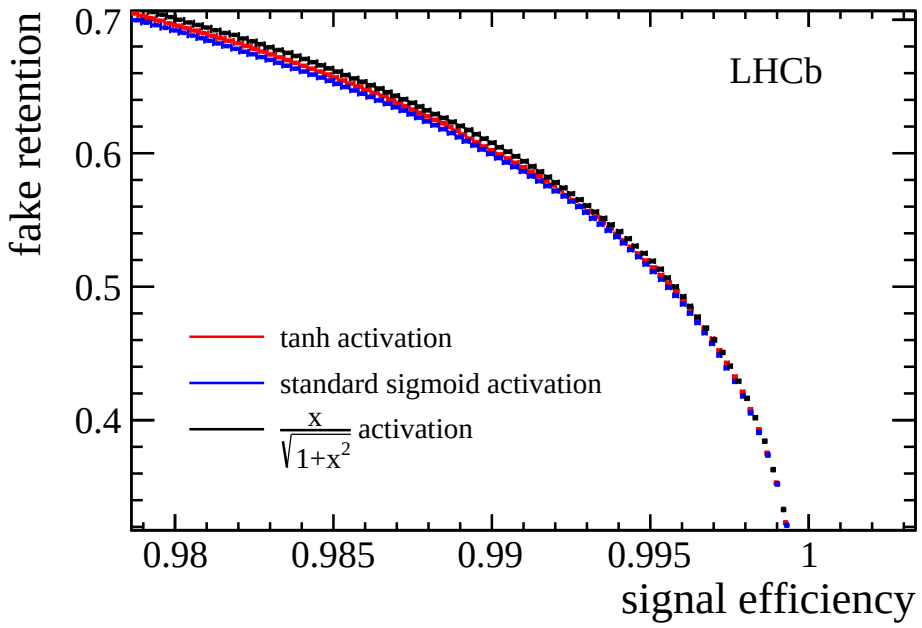

Figure 4: ROC curves for fake track discriminating neural networks, using different activation functions. A very small physics performance improvement is observed when changing from the TMVA standard functions to $\frac{x}{\sqrt{1+x^2}}$.

Of the tested functions $\frac{x}{\sqrt{1+x^2}}$ is the fastest to compute, while no significant physics performance difference is expected given the similar functional shape, see Fig. 3. No performance difference is observed in Fig. 4. Therefore, it is chosen as activation function.

The ghost probability is a classification problem, and thus cross entropy [9] is chosen as loss function in the network training. With respect to the run I implementation of the ghost probability, this contributes to the physics performance improvement. Technically the output layer activation function when training with cross entropy loss must be $\frac{1}{1-e^{-x}}$ in TMVA, it is removed from the network after the training and replaced by a custom calibration as described in Sect. 1.6).

1.5 Variable selection

To allow for enough development time for testing and evaluation, the selection of input variables is mostly unchanged from Run I with two exceptions. The track interpolation to determine the number of expected hits is removed to reduce the CPU usage of the ghost probability by a factor 10. The number of track candidates competing for shared hits in the pattern recognition added as input variable.

1.6 Output transformation

To ease the usage of the ghost probability, a transformation of the network is applied. A probability integral transform[2] is obtained as a linear spline fit to the cumulative network response for fake tracks in simulated events. The discriminating behaviour of any classifier is invariant under monotonous transformations and so is the physics performance under the probability integral transform. Motivations for this transformation are primarily to give a physical interpretation to the response: rejecting tracks with a ghost probability of larger than $x\,\%$ will retain $x\,\%$ of all fake tracks.

In addition will any update of the ghost probability training have the same behaviour and thus optimal working points of algorithms downstream of the ghost probability algorithm will remain unchanged at leading order.

1.7 Category classifiers

Fake tracks produced by different pattern recognition algorithms might have different track properties. It might therefore be beneficial to train separate neural networks for the two main track reconstruction algorithms at LHCb. On simulated events, the physics performance of two separate networks does not differ from the physics performance of a single network. Similarly, different networks for different T station regions have been tested (one for tracks in the OT, IT, and the overlap region), without significant performance gain.

Consequently, a single network for all pattern recognition algorithms in the entire detector is deployed.

[2] also referred to as "flattening" or "rarity transformation"

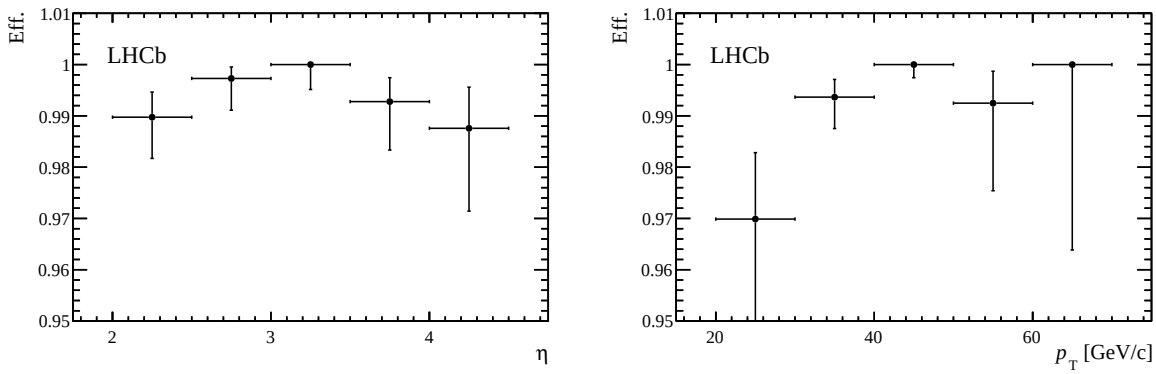

Figure 5: Efficiency for $Z \to \mu\mu$ tracks to pass the ghost probability, in data from 2015 at a bunch spacing of 50 ns.

2 Validation

The data taking strategy of LHCb in Run II involves the application of the same track reconstruction in the software trigger as in the offline data processing. This goal can only be achieved by applying the ghost probability in the software trigger to reduce the rate of fake tracks entering the particle identification and combinatorics of decay reconstructions.

It must therefore be ensured that the full physics program of LHCb can done with tracks passing the ghost probability, and that there is no corner of phase space or particle species, which is rejected by the ghost probability.

The latest update to the track fit χ^2 computation, as used since 25 ns data taking in 2015, was applied in all validations by refitting all tracks prior to computing the ghost probability.

2.1 High momentum tracks (data from 2015, 50 ns bunch spacing)

In the early measurement period in 2015 at a bunch spacing of 50 ns, the nominal 2015 pattern recognition was used without application of the ghost probability. Refitting the candidate tracks from $Z \to \mu\mu$ decays in that period allows to assess the performance of the ghost probability for very high momentum tracks, which are absent in the training data. The measured efficiency in Fig. 5 shows that the absence of very high momentum tracks does not lead to a low efficiency.

2.2 Electron reconstruction (data from run I, 50 ns bunch spacing)

Electrons are more challenging to reconstruct than the standard candles. At the same time, it can be expected that the response of the ghost probability for electrons differs

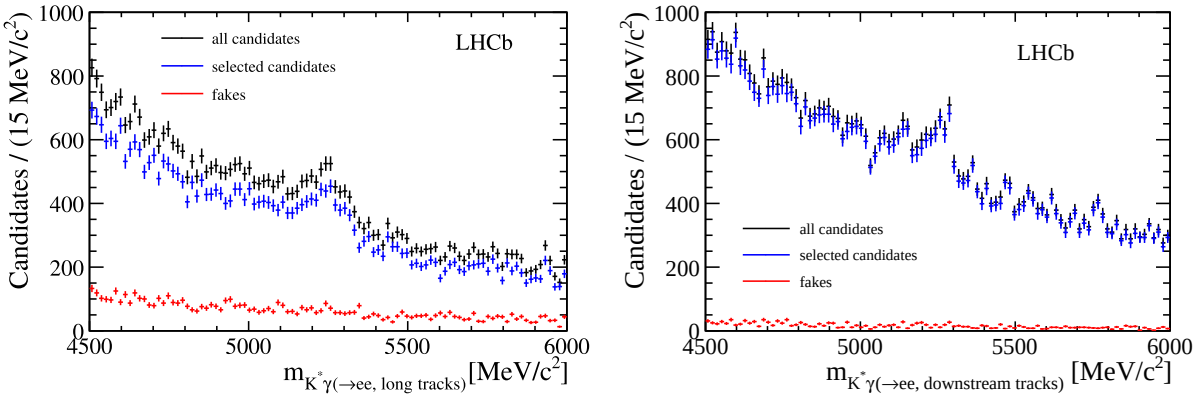

Figure 6: $B_s^0 \to K^*\gamma$ candidates, where the photon is reconstructed as pair of electron tracks, using long tracks (left) and using downstream tracks (right). The candidates prior to a cut on the ghost probability of the electron tracks are shown in black, those passing the ghost probability in blue, and those candidates rejected by the ghost probability in red. No signal loss is visible in the rejected candidates.

from that for other particles as electrons undergo more multiple scattering.

It is assumed that the reconstruction of converted photons as electron pair is the most vulnerable channel since the photon conversion can happen "late" in the VELO leaving only few hits, the e^+e^- pair has a small opening angle which could lead to hit ambiguities in the VELO pattern reconstruction, and analyses of channels $B_s^0 \to K^*\gamma$ are anyhow so-called rare decays which immediately suffer from efficiency loss.

The "early data" of 2015 does not correspond to enough integrated luminosity to obtain a satisfying estimation of the consequences of a cut on the ghost probability. The tracks of $B_s^0 \to K^*\gamma$ candidates from Run I[3] are refitted using the track fit configuration as used in 2015 and the ghost probability is evaluated. The invariant mass spectrum shown in Fig. 6 shows candidates without the application of a cut on the ghost probability, those passing, and those failing; both for converted photons reconstructed as pair of downstream tracks and as pair of long tracks. To the statistical precision of this test, no signal loss is visible.

2.3 Validation with 25 ns data from 2015

A cut on the ghost probability is included in the standard track reconstruction since data taking at 25 ns bunch spacing started. To investigate the behaviour of the ghost probability in real data with 25 ns bunch spacing, events are re-reconstructed without a cut on the ghost probability. Under the assumption, that most K^0_S are part of the underlying event and most triggered events containing K^0_S would have been triggered without those K^0_S, K^0_S are used as probe of the ghost probability.

The invariant mass spectrum of K^0_S candidates after re-reconstruction is shown in Fig. 7 for both long tracks and downstream tracks. In either case, the background contribution

[3]using a simplified version of the selection presented in [10] – without Bremsstrahlungs correction

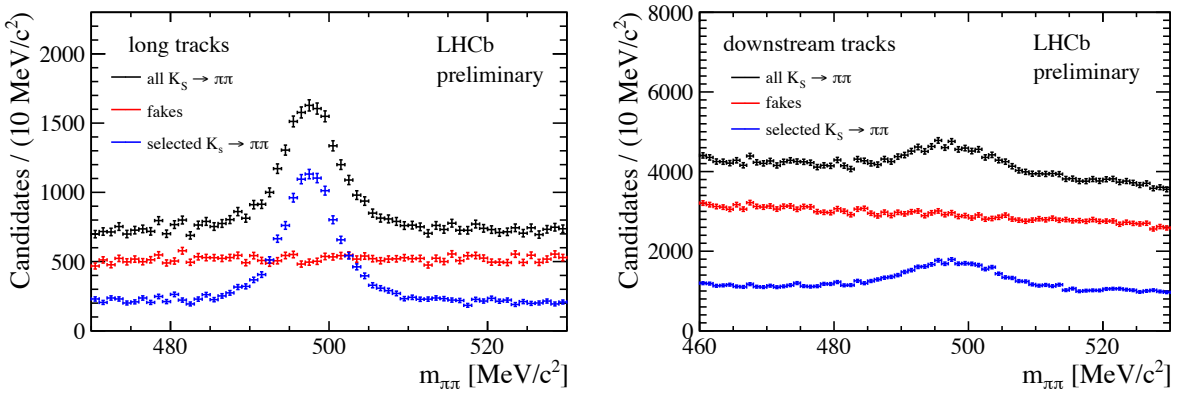

Figure 7: `StdLooseKS` invariant mass spectrum for events reconstructed without using the ghost probability in the track selection; using long tracks (left) and downstream tracks (right). Candidates for which at least one track fails the default ghost probability requirement are shown in red and do not exhibit a signal contamination. The remaining candidates are shown in blue.

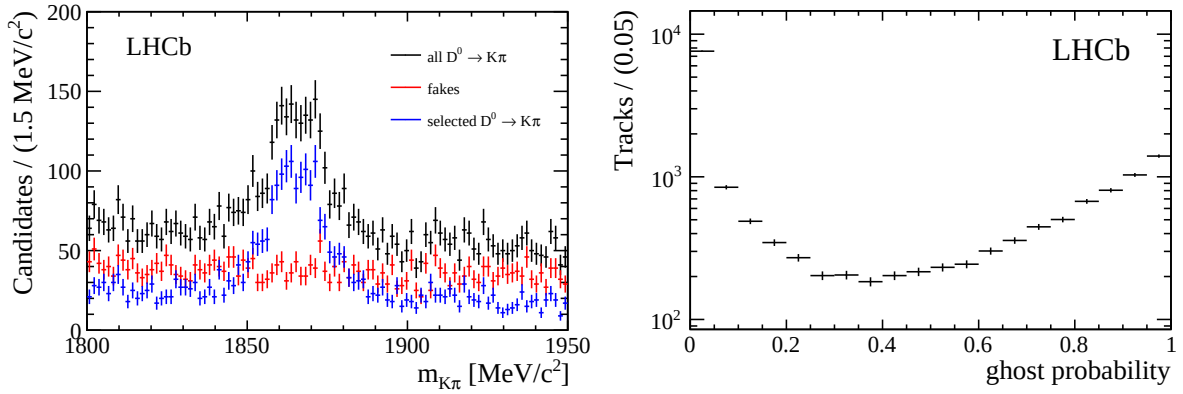

Figure 8: `StdLooseD02KPi` invariant mass spectrum for events reconstructed without using the ghost probability in the track selection (left). Candidates for which at least one track fails the default ghost probability requirement are shown in red and do not exhibit a signal contamination. The remaining candidates are shown in blue. The ghost probability distribution for the kaon and pion tracks is shown on the right.

is largely reduced when rejecting events where at least one of the tracks has a ghost probability of larger than 0.4. There is no signal visible in the events rejected. It is concluded that no physics signal is lost due to the application of the ghost probability.[4]

The same test with $D \to K\pi$ decays is shown in Fig. 8. To ensure that the sample is not biased towards candidates passing the ghost probability due to the online event selection, the ghost probability spectrum is shown in Fig. 8 b), where no step from such a selection is visible at 0.4.

[4]within the statistical sensitivity of the test; in the kinematic spectrum of the selected K^0_S

2.4 Decay time acceptance (simulated events)

The study of long lived particles (b and c hadrons) is the major part of the LHCb physics program. It must therefore be ensured that the ghost probability does not reject particles from displaced vertices at a higher probability than particles from primary collisions (which have a higher prevalence in the training).

To evaluate a possible decay time bias of the ghost probability, for each reconstructed particle in simulated events with a $b\bar{b}$ production, the average of the true decay time of their ancestor particles is determined. When rejecting tracks which fail the ghost probability criterion, the average ancestor decay time changes by $(1.5 \pm 2.0) \times 10^{-15}$ s. This is smaller than the statistical sensitivity of this test, and smaller than the systematic uncertainty to which the lifetime bias of the LHCb reconstruction is known [11].

3 Outlook

The current networks are trained for the track reconstruction for data taking in 2015 at 25 ns bunch spacing, using the latest simulations available at the time. Retrainings are advisable for significant updates in the track reconstruction "upstream" of the ghost probability (i.e. the pattern recognition and the track fit). Physics performance gains can also be expected with improved machine learning techniques or event simulations.

Additional separation between "good" tracks and fake tracks could be gained by using hit expectations in active layers: At the moment only the numbers of hits in the individual subdetectors on the track are used, these could be compared with the intersections of the trajectory with active detector material such that the number of missing hits is used as input for the network [12].

The current training is purely based on simulated events, the domain adaptation approach from [13] is not applied as it currently does not lead to an improved fake track rejection. The ghost probability network is retrained using good tracks and fake tracks from simulated events and unlabelled tracks from real events with the Caffe software framework [14], which is used in [13]. In addition to the network with domain adaptation, a conventional network is trained to disentangle effects from the training algorithm (TMVA $\rightarrow$ Caffe) and network architecture (adding a gradient reversal layer and domain classifier). This working point of the network responses is chosen to retain the same number of K^0_S candidates as the application of the nominal ghost probability. From the invariant mass distribution in Fig. 9, it can be seen that the TMVA network and the two networks trained in Caffe yield close to identical physics performance; the data points of the network with domain adaptation are almost entirely covered by the data points of the Caffe network without domain adaptation. This does not rule out that domain adaptation can not improve the physics performance of the ghost probability in the future.

The current network relies on auto-vectorisation. The methods suggested by [15] lead to tenfold improvement of the neural network implementation [16] once using AVX intrinsic commands. This approach has not been followed up to ensure platform independence of the ghost probability.

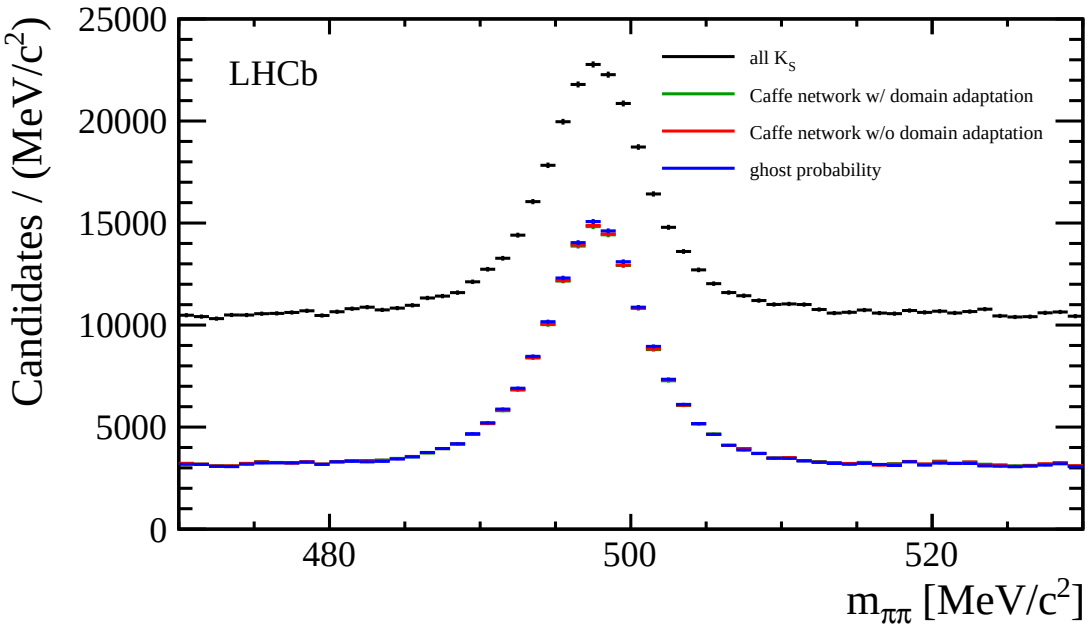

Figure 9: Comparison of the K^0_S selection with the nominal ghost probability (blue), a network trained with the Caffe package (red), and a network trained with domain adaptation (green, covered by red). The working points of the Caffe networks are chosen to retain the same numbers of candidates. Hardly any performance difference is visible.

Table 1: Callgrind benchmark comparisons of different activation functions. Fields with n/a have not been evaluated or are not available with AVX intrinsics. The activation function used by the ghost probability is marked with (*).

function	default compiler options	AVX vectorisation by hand
tanh	19,355,124,355	n/a
$\frac{1}{1+e^{-x}}$	21,140,125,632	n/a
$\frac{x}{\sqrt{1+x^2}}$(*)	415,121,741	195,121,939
$\frac{x}{1+\lvert x\rvert}$	395,121,798	195,104,759
$\max(0, x)$	155,095,875	115,095,891

The current activation function in the neural network is $\frac{x}{\sqrt{1+x^2}}$. The rectified linear unit $\max(0, x)$ or $\frac{x}{1+|x|}$ are expected to be even faster, as listed in Tab. 1 (from [17]).

4 Conclusion

The ghost probability is a super awesome tool which allows LHCb to do super good physics, and we know we can use it for everything and it makes our software trigger fast, which means we can do more fancy stuff with the CPU resources.

References

[1] J. Weidendorfer, M. Kowarschik, and C. Trinitis, *A Tool Suite for Simulation Based Analysis of Memory Access Behavior*, in *Computational Science - ICCS 2004, 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part III* (M. Bubak, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, eds.), vol. 3038 of *Lecture Notes in Computer Science*, pp. 440–447, Springer, 2004. http://www.valgrind.org/docs/pubs.html.

[2] LHCb collaboration, R. Aaij *et al.*, *Measurement of the track reconstruction efficiency at LHCb*, JINST **10** (2015) P02007, arXiv:1408.1251.

[3] J. Brehmer, J. Albrecht, and P. Seyfert, *Ghost probability: an efficient tool to remove background tracks*, https://cds.cern.ch/record/1478372. LHCb internal note LHCb-INT-2012-025.

[4] LHCb collaboration, R. Aaij *et al.*, *Observation of $B_c^+ \to J/\psi D_s^+$ and $B_c^+ \to J/\psi D_s^{*+}$ decays*, Phys. Rev. **D87** (2013) 112012, arXiv:1304.4530.

[5] A. Hoecker *et al.*, *TMVA: Toolkit for Multivariate Data Analysis*, PoS **ACAT** (2007) 040, arXiv:physics/0703039.

[6] M. Schiller, H. Voss, and L. Moneta, *fast tanh implementation*, ROOT-7054.

[7] Wikipedia. Sigmoid function, 2015.

[8] P. Seyfert and H. Voss, *TActivation... implementations*, ROOT-7062.

[9] C. M. Bishop, *Pattern recognition and machine learning*, Information science and statistics, Springer, New York [u.a.], 10. (corrected at 8th printing) ed., 2009; J.-H. Zhong *et al.*, *A program for the Bayesian Neural Network in the ROOT framework*, Comput. Phys. Commun. **182** (2011) 2655, arXiv:1103.2854.

[10] L. Beaucourt, E. Tournfier, M. N. Minard, and J. F. Marchand, *$B^0 \to K^*\gamma(e^+e^-)$ and $B_s^0 \to \phi\gamma(e^+e^-)$ analysis status*, in *2nd Radiative decays @LHCb Workshop* (A. Oyanguren Campos *et al.*, eds.), 2015. https://indico.cern.ch/event/375424/.

[11] LHCb collaboration, R. Aaij *et al.*, *Measurements of the B^+, B^0, B_s^0 meson and Λ_b^0 baryon lifetimes*, JHEP **04** (2014) 114, arXiv:1402.2554.

[12] W. Hulsbergen, *LHCBPS-1414*, https://its.cern.ch/jira/browse/LHCBPS-1414.

[13] Y. Ganin and V. Lempitsky, *Unsupervised Domain Adaptation by Backpropagation*, ArXiv e-prints (2014) arXiv:1409.7495.

[14] Y. Jia *et al.*, *Caffe: Convolutional Architecture for Fast Feature Embedding*, arXiv:1408.5093.

[15] V. Vanhoucke, A. Senior, and M. Z. Mao, *Improving the speed of neural networks on CPUs*, https://research.google.com/pubs/archive/37631.pdf.

[16] P. Seyfert, *github project* `TMVA-MLP`, https://github.com/pseyfert/tmva-mlp.

[17] P. Seyfert, *CPU performance comparison of activation functions*, https://twiki.cern.ch/twiki/bin/view/Main/PaulSeyfert?forceShow=1#activation_function.