Reducing Systematic Bias in Machine Learning Applications to J/ψ Signal Extraction in High-Energy Nuclear Physics

Authors: Yan Wang, Rangrong Ma, Kaifeng Shen, Zebo Tang, Wangmei Zha, Zebo Tang, Wangmei Zha

Date: 2026-03-30T17:13:07+00:00

Abstract

Machine-learning techniques are increasingly used in high-energy nuclear physics because they can exploit multivariate correlations more efficiently than conventional cut-based analyses. A central challenge is the construction of training samples that faithfully reproduce the detector response observed in data. Signal samples are usually derived from detector simulations; therefore, mismatches between simulation and data can degrade classifier performance and introduce systematic biases. This work presents two practical correction procedures, namely cumulative distribution function (CDF) mapping and a shift-and-scale transformation, to align simulated signal features with those measured in data. Their performance is demonstrated with J/ψ yield measurements in √sNN = 200 GeV Ru+Ru and Zr+Zr collisions recorded by STAR. A set of self-consistency tests shows that these procedures substantially suppress the systematic bias associated with data-simulation discrepancies in machine-learning-based signal extraction.

Full Text

Preamble

Reducing Systematic Bias in Machine Learning Applications to J/ψ Signal Extraction in High-Energy Nuclear Physics∗

Yan Wang,1 Rangrong Ma,2 Kaifeng Shen,1 Zebo Tang,1, † and Wangmei Zha1, ‡ 1Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China 2Physics Department, Brookhaven National Laboratory, Upton, New York 11973, USA

Machine-learning techniques are increasingly used in high-energy nuclear physics because they can exploit multivariate correlations more efficiently than conventional cut-based analyses. A central challenge is the con- struction of training samples that faithfully reproduce the detector response observed in data. Signal samples are usually derived from detector simulations; therefore, mismatches between simulation and data can degrade classifier performance and introduce systematic biases. This work presents two practical correction procedures, namely cumulative distribution function (CDF) mapping and a shift-and-scale transformation, to align simulated signal features with those measured in data. Their performance is demonstrated with J/ψ yield measurements sNN = 200 GeV Ru+Ru and Zr+Zr collisions recorded by STAR. A set of self-consistency tests shows that these procedures substantially suppress the systematic bias associated with data-simulation discrepancies in machine-learning-based signal extraction.

Keywords

Machine learning, J/ψ reconstruction, cumulative distribution function mapping, shift-and-scale

INTRODUCTION

Machine learning has become an indispensable tool across a broad range of scientific and technological disciplines, in- cluding image recognition, natural-language processing, rec- ommendation systems, and speech analysis.

Its main ad- vantage is the ability to identify complex structures in high- dimensional datasets and to provide powerful tools for classi- fication, regression, and clustering [1–6]. In nuclear physics, machine learning has been used to extract or predict key ob- servables such as charge radii, nuclear masses, and binding energies [7–11]. In relativistic heavy-ion collisions, it is es- pecially valuable for classification tasks, for example sep- arating rare physics signals from large combinatorial back- grounds [12–17]. Representative algorithms include boosted decision trees (BDTs) [18, 19], deep neural networks [20–22], and ensemble methods [23], all of which can exploit mul- tidimensional correlations more effectively than conventional straight-cut analyses. Independent of the specific algorithm, a reliable classifier requires representative signal and background samples. Whereas background samples can often be obtained directly from data, signal samples are usually pro- duced with Monte Carlo simulations. Because detector sim- ulations do not perfectly reproduce the data, training on mis- matched signal samples can degrade classifier performance and bias the extracted physics observables, as illustrated in Appendix A.

To address this challenge, we propose two correction meth- ods: cumulative distribution function (CDF) mapping and shift-and-scale. CDF mapping systematically transforms a simulated feature distribution into the corresponding distribu- tion measured in data without requiring an explicit parametric

∗ Supported by National Key R&D Program of China under grant No. 2024YFA1610803 and the National Natural Science Foundation of China under grant Nos. 12361141827 and 12175223.

† Corresponding author: zbtang@ustc.edu.cn ‡ Corresponding author: first@ustc.edu.cn

description of the shape, which makes it flexible and broadly applicable. The shift-and-scale procedure is more restrictive but simpler to implement, and it is well suited to cases in which the simulated and measured distributions already have similar shapes and differ mainly in location and width.

J/ψ is a classic probe of the medium created in relativis- tic heavy-ion collisions and is widely used to investigate the properties of the Quark–Gluon Plasma [24, 25]. Accurate yield measurements are therefore of continuing interest [26–In this work, a BDT implemented with the eXtreme Gradient Boosting (XGBoost) framework [35] is used to sep- arate the J/ψ signal from background in Ru+Ru and Zr+Zr sNN = 200 GeV recorded by STAR at RHIC. collisions at This case study demonstrates both the implementation of the proposed correction procedures and their effectiveness in im- proving the robustness of machine-learning-based signal ex- traction. Section II describes the J/ψ reconstruction, Sections III and IV present the construction of the training samples and the model training procedure, Section V discusses the self- consistency tests and results, and Section VI summarizes the main conclusions.

II. J/ψ RECONSTRUCTION

In 200 GeV Ru+Ru and Zr+Zr collisions, J/ψ mesons are reconstructed through the dielectron decay channel [27, 34, 36], with electron4 identification based on the Time-Of-Flight (TOF) [37], the Time Projection Chamber (TPC) [38, 39], and the Barrel Electromagnetic Calorimeter (BEMC) [40] detectors. Loose preselection requirements on the particle- identification (PID) observables are first imposed to suppress the dominant background while retaining sufficient signal statistics for the subsequent machine-learning analysis.

4 Here, “electron” denotes both electrons and positrons unless otherwise

specified.

The TOF detector measures the arrival time of a charged particle. Together with the collision start time and the mo- mentum measured in the TPC, one obtains $1/\beta$ c/v, where c is the speed of light and v is the particle velocity. Figure 1 [Figure 1: see original paper] shows the distribution of $1/\beta$ as a function of momentum p for charged particles. Because particles with different masses have different velocities at the same momentum, distinct PID bands appear in this representation. A preselection cut of $|1/\beta - 1| < 0.035$ is applied, shown as dashed lines in Fig. 1 and listed in Tab. 1.

after applying the $|1/\beta - 1| < 0.025$ cut. The dashed curves corre- spond to the nσe cuts used for electron preselection.

where the dashed horizontal lines indicate the $1/\beta$ cuts for electron preselection.

The TPC measures the specific ionization energy loss, dE/dx, which provides additional separation between elec- trons and hadrons. PID based on dE/dx is commonly ex- pressed through the normalized deviation of the measured value from the expectation for a given particle species, de- noted by nσe, n$\sigma\pi$, and so forth. For electrons, the nσe dis- tribution is expected to be approximately Gaussian with mean 0 and width 1 after proper calibration. Figure 2 [Figure 2: see original paper] shows the nσe distribution as a function of p after imposing a tighter TOF re- quirement, $|1/\beta - 1| < 0.025$, to improve the electron purity for visualization. The dashed curves indicate the momentum- dependent preselection criteria used before machine-learning training.

At high momentum, the BEMC provides additional PID through the ratio of the deposited electromagnetic energy to the track momentum. In practice, electron identification com- monly uses the largest tower energy within the BEMC clus- ter, E0, rather than the total cluster energy Etot, because E0 has better resolution and is less sensitive to the high detec- tor occupancy in heavy-ion collisions. Figure 3 [Figure 3: see original paper] compares the E0/p distributions for protons and electrons in the range $4 < \mathrm{pT} < 6$ GeV/c. The electron sample, often referred to as photonic electrons, is obtained from photon conversions and neutral-meson Dalitz decays [41]. As expected, the electron distribution peaks at a larger E0/p value than the proton dis- tribution. The dashed vertical lines indicate the preselection requirement on E0/p.

The use of the TOF, TPC, and BEMC preselection crite-

(solid line) within $4 < \mathrm{pT} < 6$ GeV/c. The vertical dashed lines correspond to the E0/p cuts used for electron preselection.

ria depends on the track kinematics and on whether matching information is available in the TOF and BEMC, as summa- rized in Tab. 1. Electron and positron candidates that sat- isfy these requirements are paired to form J/ψ candidates, and the resulting invariant-mass distribution is shown in Fig. 4 Figure 4: see original paper.

The combinatorial background is estimated with the mixed- event technique [42]. After subtracting this component, the remaining invariant-mass distribution is fitted with two terms: a Crystal Ball function for the J/ψ signal and a linear term for the residual background arising mainly from semilep- tonic heavy-flavor decays and Drell–Yan production. The raw J/ψ yield is obtained by counting the electron–positron pairs in the mass window [2.91, 3.21] GeV/c2 and subtract- ing the fitted residual-background contribution in the same interval. Approximately 1.1×105 J/ψ candidates are recon- structed, with a signal significance of 103. The following sec- tions show how machine learning further improves this significance.

00.511.522.53p (GeV/c)0.70.80.911.11.21.31.41.51.6b1/110210310410510610710810empkp
= 200 GeVNNsRu+Ru & Zr+Zr, 00.511.522.533.54p (GeV/c)10-5-0510esn1102103104105106107 10epkp

= 200 GeVNNsRu+Ru & Zr+Zr, 00.511.522.533.54/p0E020406080100120140310• CountsRu+Ru & Zr+Zr, 200 GeV: 4 - 6 GeV/cTpElectronProton

Track pT

Detectors used

pT $\leq$ 1.0 GeV/c

TPC, TOF

pT > 1.0 GeV/c

TPC, TOF and BEMC no matching

TPC, BEMC and TOF no matching

TPC, TOF and

Electron preselection cuts $|1/\beta - 1| < 0.035$; for p > 0.8 GeV/c: $-1 < n\sigma e < 2.25$; for p $\leq$ 0.8 GeV/c: $(3 \times p - 3.4) < n\sigma e < 2.25$; $|1/\beta - 1| < 0.035$; $-1 < n\sigma e < 2.25$

$-1.25 < n\sigma e < 2.25$; $0.3 < E0/p < 1.8$

$|1/\beta - 1| < 0.035$; $-1.75 < n\sigma e < 2.25$; $0.3 < E0/p < 1.8$

BDT criterion

BDT > 0.72

BDT > 0.72

(a) With preselection cuts listed in Tab. 1

(b) With preselection and BDT cuts listed in Tab. 1

(c) With straight cuts listed in Tab. 2

Ru+Ru and Zr+Zr collisions. Different panels correspond to different electron identification cuts.

III. CONSTRUCTION OF TRAINING SAMPLES

measurement.

For J/ψ reconstruction,

the classification targets are electron–positron pairs that form J/ψ candidates. To deter- mine whether a candidate is signal-like or background-like, a set of discriminating features must be defined. In princi- ple, any observable that improves the separation without in- troducing physics bias can be used. In this work, the selected features are nσe, n$\sigma\pi$, E0/p, and the distance of closest ap- proach (DCA, defined as the minimum distance between a charged-particle trajectory and the collision vertex) for both daughter tracks, labeled as "tr1" and "tr2" .

The background training sample is selected from electron–positron pairs in the invariant-mass sidebands, namely (2.65, 2.85), (3.25, 3.5), and (3.8, 4.5) GeV/c2, well sepa- rated from the J/ψ peak. The interval (3.5, 3.8) GeV/c2 is excluded because it contains a small contribution from ψ(2S) mesons. This sideband method assumes that the background in the sidebands is representative of the background beneath the J/ψ peak. Therefore, the features used for classification should not depend strongly on the pair invariant mass, which constitutes an important feature-selection criterion. Figure 5 [Figure 5: see original paper] shows the correlations between the selected features and the invariant mass for the background sample; no strong mass de- pendence is observed. As expected, nσe and n$\sigma\pi$ are strongly correlated because both are derived from the same dE/dx

The signal training sample is generated with Monte Carlo simulation. Electrons and positrons from J/ψ decays are prop- agated through a GEANT3-based [43] model of the STAR de- tector, and the simulated hits are embedded into real Ru+Ru and Zr+Zr events before reconstruction with the same soft- Inevitably, the simulation does ware chain used for data. not perfectly reproduce the data because detector conditions evolve with time and the heavy-ion environment has high oc- cupancy. The quality of the simulation can be assessed by comparing the nσe, n$\sigma\pi$, E0/p, and DCA distributions for electrons from simulated J/ψ decays with those for photonic electrons in data, as shown in Fig. 6 [Figure 6: see original paper]. The DCA distributions agree reasonably well, whereas the nσe and n$\sigma\pi$ distributions exhibit clear offsets and the E0/p distribution shows a notice- able shape mismatch. If left uncorrected, these discrepancies can degrade the machine-learning performance and bias the extracted J/ψ yield.

Because the simulated and measured nσe and n$\sigma\pi$ distribu- tions have similar shapes, they can be corrected with a shift- and-scale transformation:

nσi,cor = (nσi − µsimu) · (σdata/σsimu) + µdata,

Here, µdata (σdata) and µsimu (σsimu) are the means (stan- dard deviations) of the corresponding nσi distributions in data and simulation, respectively, with i = e or π. After this trans-

2.833.23.43.6)2 (GeV/ceeM01000200030004000500060007000310 ·]-1)2 [(GeV/ceedN/dMSignal candidatesMixed-event bkg.Bkg. subtractedCombined fityJ/Residual bkg.Ru+Ru & Zr+Zr 200 GeVCentrality: 0-80% > 0.2 GeV/cTp = 103S+BS, |y| < 1-e+ efi yJ/ 1582-= 109152 yJ/N 2.833.23.43.6)2 (GeV/ceeM020004000600080010001200310 ·]-1)2 [(GeV/ceedN/dMSignal candidatesMixed-event bkg.Bkg. subtractedCombined fityJ/Residual bkg.Ru+Ru & Zr+Zr 200 GeVCentrality: 0-80% > 0.2 GeV/cTp = 173S+BS, |y| < 1-e+ efi yJ/ 614-= 76010 yJ/N 2.833.23.43.6)2 (GeV/ceeM020004000600080010001200310 ·]-1)2 [(GeV/ceedN/dMSignal candidatesMixed-event bkg.Bkg. subtractedCombined fityJ/Residual bkg.Ru+Ru & Zr+Zr 200 GeVCentrality: 0-80% > 0.2 GeV/cTp = 139S+BS, |y| < 1-e+ efi yJ/ 670-= 65647 yJ/N 4

the simulation and data differ strongly in shape, this discrete matching can produce bin-to-bin fluctuations at the level of one histogram bin, leading to localized excesses in some bins and deficits in neighboring bins. This effect explains the sta- tistical irregularities visible in Fig. 6(c). Because the classifier is trained on event-by-event continuous feature values rather than on the displayed histogram counts themselves, this bin- ning artifact does not compromise the model training. It can be further reduced by decreasing the histogram bin width.

Compared with shift-and-scale, CDF mapping is more gen- eral but also more involved to implement. Both procedures act deterministically on each event and therefore retain the event-level correspondence among observables, which is essential for multivariate classifiers to benefit from joint feature information. For the same reason, the signal sample cannot be constructed by independently sampling one-dimensional fea- ture distributions from data, because that procedure would de- stroy the physically relevant inter-feature dependence struc- ture.

IV. MODEL TRAINING AND APPLICATION

A BDT classifier implemented with XGBoost is trained to further improve the J/ψ signal significance. The signal and background samples are each divided into two equal subsets, one for training and the other for testing. After training, the model is applied to both subsets. For every electron–positron pair, the classifier returns an output between 0 and 1, with val- ues close to 1 corresponding to signal-like candidates and val- ues close to 0 corresponding to background-like candidates.

The output distributions are shown in Fig. 7 [Figure 7: see original paper]. The red band and red points denote the signal distributions for the training and testing samples, respectively, whereas the blue band and blue points show the corresponding background distributions.

The strong separation between signal and background indi- cates good discriminating power, and the agreement between the training and testing samples shows that overtraining is not evident.

In the final analysis, a threshold on the BDT output is cho- sen so that pairs with scores above the threshold are retained as signal candidates and those below it are rejected as back- ground. For a given threshold, the signal efficiency is de- fined as the fraction of signal events that survive the cut, and the background efficiency is defined analogously. Plotting the signal efficiency against the background efficiency while scanning the threshold from 0 to 1 yields the receiver oper- ating characteristic (ROC) curve, where the horizontal and vertical axes correspond to the false-positive rate (FPR) and true-positive rate (TPR), respectively. The ROC curves for the training and testing samples are shown in Fig. 8 [Figure 8: see original paper]. The area under the curve (AUC) quantifies the overall classifica- tion performance; values closer to 1 indicate better separa- tion. The AUC obtained here is 0.91, demonstrating strong discriminating power. The nearly identical ROC curves and AUC values for the training and testing samples provide ad- ditional evidence that the model is not overfit.

structed electron–positron pair invariant mass for the background training sample. Each cell represents the correlation coefficient be- tween a pair of variables, with positive correlations shown in red and negative correlations in blue.

formation, the simulated nσe and n$\sigma\pi$ distributions agree well with those observed in data, as shown in Fig. 6(a) and

For the E0/p distribution, the mismatch between simula- tion and data is not limited to a simple offset or width differ- ence, so shift-and-scale is insufficient. We therefore introduce a cumulative distribution function (CDF) mapping method.

Let fsimu(x) and fdata(x), with x [xmin, xmax], denote the probability density functions of a feature in simulation and data, respectively. The corresponding CDF is

C(x) =

(cid:90) x

f (x′) dx′.

For an event in simulation with feature value A, its cumulative probability is

Csimu(A) =

(cid:90) A

fsimu(x′) dx′.

The corrected value B is then obtained from the data distri- bution through

Cdata(B) = Csimu(A) B = C −1

data(Csimu(A)).

This procedure defines an event-by-event mapping from the original value A to the corrected value B, such that the transformed simulation reproduces the data distribution. Af- ter CDF mapping, the corrected simulated E0/p distribution (open diamonds) is consistent with the distribution measured in data (solid circles), as shown in Fig. 6(c). In practice, the PDFs are represented by discrete histograms, so the method is implemented numerically by constructing discrete CDFs and searching for the closest cumulative probability. When

pairMasstr1E0Optr1NsigmaEtr1NsigmaPitr1Dcatr2E0Optr2NsigmaEtr2NsigmaPitr2DcapairMasstr1E0Optr1
1.00-0.75-0.50-0.250.000.250.500.751.00 5

(a) nσe distributions

(b) n$\sigma\pi$ distributions

(c) E0/p distributions

(d) DCA distributions

(open diamonds) and without (filled diamonds) corrections, to those from real data (filled circles).

V. MODEL VALIDATION VIA SELF-CONSISTENCY CHECKS

Beyond establishing that the model is not overtrained, it is necessary to verify that the BDT efficiencies for both signal and background are correctly reproduced in real data. This validation is carried out through two self-consistency tests.

The first test examines the J/ψ yield corrected for the BDT efficiency as a function of the BDT threshold. Because the measured yield should not depend on the particular analy- sis threshold, the corrected yield is expected to remain con- stant within uncertainties. The J/ψ efficiency for a given BDT threshold is derived from the signal output distribution in Fig. 7, and the resulting efficiency curves are shown in lisions. Here, centrality characterizes the geometric overlap of the two nuclei: 0–20% corresponds to the most central col- lisions, whereas 40–80% corresponds to more peripheral col-

lisions. After applying the preselection criteria in Tab. 1 and varying the BDT threshold, the raw J/ψ counts are extracted from data with the procedure described in Sec. II. Figure 10 [Figure 10: see original paper] then shows the efficiency-corrected J/ψ counts as a function of the BDT threshold for four centrality classes. Although the signal efficiency changes by roughly an order of magnitude as the threshold increases from 0 to 0.9, the corrected counts re- main stable, with variations below 2%. This stability strongly suggests that the signal-efficiency estimate from the machine- learning model accurately represents the efficiency in data, thereby validating the feature-correction procedure described in Sec. III.

A more stringent validation compares the J/ψ significance as a function of the BDT threshold between the machine- learning expectation and the measurement in data, because this test depends on both the signal and background efficien-

5-4-3-2-1-0123esn00.10.20.30.40.50.6Probability densitySimulation with correctionSimulationDataRu+Ru & Zr+Zr = 200 GeVNNsCentrality: 0-80%| < 0.6 h0.4< |: 1.2 - 1.6 GeV/c Tp 2-0246810psn00.10.20.30.40.5Probability densitySimulation with correctionSimulationDataRu+Ru & Zr+Zr = 200 GeVNNsCentrality: 0-80%| < 0.6 h0.4< |: 1.2 - 1.6 GeV/c Tp 00.511.522.533.54/p0E00.511.522.5Probability densitySimulation with correctionSimulation DataRu+Ru & Zr+Zr = 200 GeVNNsCentrality: 0-80%: 1.2 - 1.6 GeV/c Tp| < 0.6 h0.4< | 00.10.20.30.40.50.60.70.80.91Dca012345Probability densityDataSimulationRu+Ru & Zr+Zr = 200 GeVNNsCentrality: 0-80%: 1.2 - 1.6 GeV/c Tp| < 0.6 h0.4< | 6

ing samples. The red (blue) histogram represents the training signal (background) sample, while the red (blue) points are for the testing signal (background) sample.

ent centrality classes of 200 GeV Ru+Ru and Zr+Zr collisions.

ples. The reported AUC values quantify the overall performance, while the diagonal line represents a random classifier for compari-

cies. The expected J/ψ significance is calculated as

S(x) = S0 •effs(x)/effs(x0), B(x) = B0 •effb(x)/effb(x0), Significanceexpected(x) = S(x)/(cid:112)S(x) + B(x).

In Eq. 2, x is the BDT threshold, and effs(x) and effb(x) are the corresponding signal and background efficiencies. The quantities S(x) and B(x) denote the signal and background counts after the threshold is applied, while S0 and B0 are the counts at a reference threshold x0. The choice of x0 is arbi- trary and is used only to normalize the prediction to the data at one point. The resulting expected significance is shown as the gray band in Fig. 11 [Figure 11: see original paper]. It first increases and then decreases as the threshold is tightened from 0 to 0.9, reflecting the com-

of BDT criterion in different centrality classes of 200 GeV Ru+Ru and Zr+Zr collisions.

petition between improved background rejection and reduced signal efficiency. The prediction from the trained model is compared with the significance extracted directly from data, shown as open circles. For clarity, the two curves are aligned at a threshold of 0.5, namely x0 = 0.5 in Eq. 2. Their con- sistency indicates that the training samples reproduce the be- havior of both signal and background in data with sufficient accuracy for this analysis.

The expected-significance curve from the trained model also provides a principled way to determine the optimal BDT threshold. Compared with scanning the data directly, an opti- mization based on the training samples is less sensitive to sta- tistical fluctuations in the measured spectrum and therefore less prone to introducing threshold-selection bias, which is particularly important for low-statistics signals. As shown in threshold of 0.72, which is therefore chosen as the working

0.00.20.40.60.81.0Model Output10-210-1100101102103Counts (normalized)Ru+Ru & Zr+ZrpsNN=200 GeVCentrality: 080%J/ψ→e+e−, 0.2 < pT < 8.0 GeV/cBackground (train)J/ψ signal (train)Background (test)J/ψ signal (test)0.00.20.40.60.81.0False Positive Rate0.00.20.40.60.81.0True Positive RateCentrality: 080%0.2 < pT < 8.0 GeV/cTest -> ROC (AUC = 0.9093)Train -> ROC (AUC = 0.9126)Luck00.10.20.30.40.50.60.70.80.91BDT criterion00.20.40.60.81BDT efficiencyCentrality: 0-20%Centrality: 20-40%Centrality: 40-80%Centrality: 0-80% > 0.2 GeV/cTpRu+Ru & Zr+Zr = 200 GeVNNs 00.20.40.60.81BDT criterion020406080100120140160180200310• BDT Efficiency-Corrected CountsCentrality: 0-20%Centrality: 20-40%Centrality: 40-80%Centrality: 0-80%Ru+Ru & Zr+Zr = 200 GeVNNs > 0.2 GeV/cTp 7

ing J/ψ yield extraction in sNN = 200 GeV Ru+Ru and Zr+Zr collisions as a case study, we show that the corrected training samples lead to stable efficiency-corrected yields and to a predicted significance consistent with the significance measured in data. The self-consistency tests therefore vali- date both the necessity and the effectiveness of the proposed corrections. The same strategy should be broadly useful in machine-learning applications throughout high-energy nuclear and particle physics whenever simulated signal samples do not fully reproduce the data.

ACKNOWLEDGMENTS

The authors thank the STAR Collaboration and the RHIC Operations Group for providing the data used in this study.

They also acknowledge the RCF facility at Brookhaven Na- tional Laboratory for the computational resources required for this analysis. Helpful discussions with Pengzhong Lu and Zhenjun Xiong are gratefully acknowledged.

Appendix A: Results without the E0/p correction

as a function of the BDT threshold for two cases: with and without aligning the E0/p distribution of the training signal sample to that observed in data.

open circles represent the J/ψ significance directly obtained from real data while the black band shows the expected significance esti- mated from the machine learning model. The two distributions are aligned at BDT value of 0.5 as indicated by the black open square.

The vertical dashed line at 0.72 marks the point where the expected significance reaches its maximum value of 176 as indicated by the horizontal dashed line. The red star represents the J/ψ significance of 173 obtained from real data at the same BDT criterion.

point. Figure 4(b) shows the raw signal extraction obtained with this optimal threshold. The corresponding significance is 173, representing a 68% improvement over the result with only the preselection criteria [Fig. 4(a)] and a 24% improve- ment over the result obtained with straight PID cuts [Fig. 4(c) and Tab. 2].

The same self-consistency tests are also useful for diag- nosing problematic input features. If adding a given feature destroys the agreement between the model expectation and the data, that feature should be corrected or removed from the training. As an example, Fig. 12 [Figure 12: see original paper] in the Appendix shows the result obtained without CDF mapping for E0/p. In that case, the efficiency-corrected signal counts deviate progres- sively from the expected flat behavior as the BDT threshold increases, reaching 29% at a threshold of 0.9. Likewise, the expected J/ψ significance becomes inconsistent with the significance measured in data, as shown in Fig. 13 [Figure 13: see original paper]. These dis- crepancies provide direct evidence that the feature-correction procedures introduced in Sec. III are necessary for unbiased signal extraction.

VI. SUMMARY

This work develops two practical procedures, CDF map- ping and shift-and-scale, to align simulated feature distribu- tions with those measured in data while preserving the event- level multivariate information required for machine-learning classification. These methods address a common source of systematic bias in experimental analyses, namely the mis- match between simulated and real detector responses. Us-

function of BDT criterion, for scenarios that the signal sample is in- corporated with (circles) or without (diamonds) the E0/p correction.

model trained without CDF mapping for E0/p with the sig- nificance measured directly in data.

00.20.40.60.81BDT criterion050100150200250 significanceyJ/Ru+Ru & Zr+Zr 200 GeV > 0.2 GeV/cTCentrality: 0-80%, pMeasured significanceExpected significance00.20.40.60.81BDT criterion5060708090100110120130140310 • BDT Efficiency-Corrected Counts/p Correction0With E/p Correction0Without E > 0.2 GeV/cTCentrality: 0-80%, p = 200 GeVNNsRu+Ru & Zr+Zr, Appendix B: Straight electron identification cuts

from the machine learning model trained using the signal sample without the E0/p correction. It is compared to the measured signifi- cance in real data (open circles). The two distributions are aligned at the BDT criterion of 0.5, indicated by the open square, to facilitate the comparison.

The table below lists the conventional straight PID cuts used as the reference selection for comparison with the machine-learning-based approach.

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature https://doi.org/10.1038/

521, 436–444 (2015). nature14539

[2] L. PANG, K. ZHOU, X. WANG, Deep Learning for Nu- clear Physics. Nucl. Phys. Rev. 37(3), 720–726 (2020). https://doi.org/10.11804/NuclPhysRev.37.2019CNPC41

[3] F.P. Li, L.G. Pang, G.Y. Qin, QCD equation of state at finite μB using a deep learning-assisted quasi-parton model. Phys. Lett.

B 868, 139692 (2025). https://doi.org/10.1016/j. physletb.2025.139692

[4] W. He, Q. Li, Y. Ma, et al., Machine learning in nuclear physics at low and intermediate energies. Sci. China Phys. Mech.

Astron. 66(8), 282001 (2023). https://doi.org/10. 1007/s11433-023-2116-0

[5] Z. Li, Y. Wang, Q. Li, et al., Machine-learning predictions for the nuclear charge radius: Bayesian method versus decision- tree-based algorithm. Phys. Rev. C 112(1), 014312 (2025). https://doi.org/10.1103/vj25-zwd3

[6] Z.M. Niu,

J.Y. Fang, Y.F. Niu, Comparative study of radial basis function and Bayesian neural network ap- proaches in nuclear mass predictions.

Phys. Rev. C 100(5), 054311 (2019). https://doi.org/10.1103/ PhysRevC.100.054311

[7] B. Zhou, Y.Y. Cao, J.Y. Guo, Predictions of nuclear charge radii based on the convolutional neural network. Nucl. Sci.

Tech. 34(10), 152 (2023). https://doi.org/10.1007/ s41365-023-01308-x

[8] Q.F. Li, Y.J. Wang, Z.P. Gao, et al., Machine learning the nu- clear mass. Nucl. Sci. Tech. 32(10), 109 (2021). https: //doi.org/10.1007/s41365-021-00956-1

[9] L. Tang, Z.H. Zhang, Nuclear charge radius predictions by kernel ridge regression with odd-even effects. Nucl. Sci.

Tech. 35(2), 19 (2024). https://doi.org/10.1007/ s41365-024-01379-4

[10] Z.Z. Ren, Z.Y. Yuan, D. Bai, et al., Reliable calculations of nuclear binding energies by the gaussian process of machine learning. Nucl. Sci. Tech. 35(6), 105 (2024). https://doi. org/10.1007/s41365-024-01463-9

[11] G.Q. Zhang, X.G. Cao, G.Y. Cheng, et al., The study of intelligent algorithm in particle identification of heavy-ion collisions at Nucl. Sci. low and intermediate energies.

Tech. 35(2), 33 (2024). https://doi.org/10.1007/ s41365-024-01388-3

[12] K. Albertsson, et al., Machine Learning in High Energy Physics Community White Paper.

J. Phys. Conf. Ser. 1085(2), 022008 (2018). https://doi.org/10.1088/ 1742-6596/1085/2/022008

[13] A.M. Sirunyan, et al. (CMS), Identification of heavy, energetic, hadronically decaying particles using machine-learning tech- niques. JINST 15(06), P06005 (2020). https://doi.org/ 10.1088/1748-0221/15/06/P06005

[14] S. Acharya, et al. (ALICE), Measurement of beauty production sNN = via non-prompt charm hadrons in p–Pb collisions at 5.02 TeV. JHEP 11, 148 (2024). https://doi.org/10. 1007/JHEP11(2024)148

[15] S. Acharya, et al. (ALICE), First measurement of Ω0

c produc- s=13 TeV. Phys. Lett. B 846, 137625 tion in pp collisions at (2023). https://doi.org/10.1016/j.physletb.

[16] S. Acharya, et al. (ALICE), First measurement of Λ+ c produc- sNN = tion down to pT = 0 in pp and p–Pb collisions at 5.02 TeV. Phys. Rev. C 107(6), 064901 (2023). https: //doi.org/10.1103/PhysRevC.107.064901 [17] S. Acharya, et al. (ALICE), Study of flavor dependence of the baryon-to-meson ratio in proton-proton collisions at 13 TeV. Phys. Rev. D 108(11), 112003 (2023). https:

//doi.org/10.1103/PhysRevD.108.112003 [18] Y. Coadou, Boosted decision trees (2022). https://doi.

org/10.1142/9789811234033_{0002}

00.20.40.60.81BDT criterion050100150200250 significanceyJ/Ru+Ru & Zr+Zr 200 GeV > 0.2 GeV/cTCentrality: 0-80%, pMeasured significanceExpected significance Track pT

pT $\leq$ 1.0 GeV/c

pT > 1.0 GeV/c

Detectors used

TPC, TOF

TPC, TOF and BEMC no matching

TPC, BEMC and TOF no matching

TPC, TOF and

Electron identification cuts $|1/\beta - 1| < 0.025$; for p > 0.8 GeV/c: -0.75 < nσe < 2; for p $\leq$ 0.8 GeV/c: (3$\times$ p - 3.15) < nσe < 2 $|1/\beta - 1| < 0.025$; $-0.75 <$ nσe < 2

$-1 <$ nσe < 2; 0.5 < E0/p < 1.5

$|1/\beta - 1| < 0.025$; $-1.5 <$ nσe < 2; 0.5 < E0/p < 1.5

[19] B.P. Roe, H.J. Yang, J. Zhu, Y. Liu, et al., Boosted decision trees, an alternative to artificial neural networks. Nucl. Instrum.

Meth. A 543(2-3), 577–584 (2005). https://doi.org/ 10.1016/j.nima.2004.12.018

[20] J. Schmidhuber, Deep learning in neural networks: An overview. Neural Networks 61, 85–117 (2015). https: //doi.org/10.1016/j.neunet.2014.09.003 [21] C.F. Madrazo, I.H. Cacha, L.L. Iglesias, J.M. de Lucas, Ap- plication of a Convolutional Neural Network for image classi- fication for the analysis of collisions in High Energy Physics.

EPJ Web Conf. 214, 06017 (2019). https://doi.org/ 10.1051/epjconf/201921406017

[22] P. Baldi, K. Bauer, C. Eng, et al., Jet Substructure Classifi- cation in High-Energy Physics with Deep Neural Networks.

Phys. Rev. D 93(9), 094034 (2016). https://doi.org/ 10.1103/PhysRevD.93.094034

[23] I. Bentley, J. Tedder, M. Gebran, et al., High precision binding energies from physics-informed machine learning. Phys. Rev.

C 111(3), 034305 (2025). https://doi.org/10.1103/ PhysRevC.111.034305

[24] T. Matsui, H. Satz, J/ψ Suppression by Quark-Gluon Plasma Formation. Phys. Lett. B 178, 416–422 (1986). https:// doi.org/10.1016/0370-2693(86)91404-8

[25] M. Gyulassy, The QGP discovered at RHIC, in NATO Ad- vanced Study Institute: Structure and Dynamics of Elementary Matter (2004), pp. 159–182

[26] L. Adamczyk, et al. (STAR), J/ψ production at low pT in Au sN N = 200 GeV with the + Au and Cu + Cu collisions at STAR detector. Phys. Rev. C 90(2), 024906 (2014). https: //doi.org/10.1103/PhysRevC.90.024906

[27] L. Adamczyk, et al. (STAR), J/ψ production at high transverse sN N = 200 GeV. momenta in p+p and Au+Au collisions at Phys. Lett. B 722, 55–62 (2013). https://doi.org/10. 1016/j.physletb.2013.04.010

[28] L. Adamczyk, et al. (STAR), Energy dependence of J/ψ pro- sN N = 39, 62.4 and 200 duction in Au+Au collisions at GeV. Phys. Lett. B 771, 13–20 (2017). https://doi. org/10.1016/j.physletb.2017.04.078

[29] A. Adare, et al. (PHENIX), J/ψ Production vs Central- ity, Transverse Momentum, and Rapidity in Au+Au Colli- sN N = 200 GeV. Phys. Rev. Lett. 98, 232301 sions at (2007). https://doi.org/10.1103/PhysRevLett.

[30] B.B. Abelev, et al. (ALICE), Centrality, rapidity and trans- verse momentum dependence of J/ψ suppression in Pb-Pb sNN=2.76 TeV. Phys. Lett. B 734, 314–327 collisions at (2014). https://doi.org/10.1016/j.physletb.

[31] J. Adam, et al. (ALICE), Inclusive, prompt and non-prompt J/ψ production at mid-rapidity in Pb-Pb collisions at sNN = 2.76 TeV. JHEP 07, 051 (2015). https://doi.org/10. 1007/JHEP07(2015)051

[32] J. Adam, et al. (STAR), Measurement of inclusive J/ψ sup- sN N = 200 GeV through pression in Au+Au collisions at the dimuon channel at STAR. Phys. Lett. B 797, 134917 (2019). https://doi.org/10.1016/j.physletb.

[33] Y. Wang (STAR), J/ψ production in Ru + Ru and Zr + Zr sNN = 200 GeV with the STAR experiment. collisions at PoS HardProbes2023, 110 (2024). https://doi.org/ 10.22323/1.438.0110

[34] Z. Tang, W. Zha, Y. Zhang, An experimental review of open heavy flavor and quarkonium production at RHIC. Nucl. Sci.

Tech. 31(8), 81 (2020). https://doi.org/10.1007/ s41365-020-00785-8

[35] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting Sys- tem (2016). https://doi.org/10.1145/2939672.

[36] J. Chen, et al., Properties of the QCD matter:

review of selected results from the relativistic heavy ion collider beam energy scan (RHIC BES) program. Nucl. Sci. Tech. https://doi.org/10.1007/ 35(12), 214 (2024). s41365-024-01591-2

[37] W.J. Llope (STAR), The large-area time-of-flight (TOF) up- grade for the STAR detector. AIP Conf. Proc. 1099(1), 778-781 (2009). https://doi.org/10.1063/1.3120153 [38] H. Wieman, et al. (STAR), STAR TPC at RHIC. IEEE Trans.

Nucl. Sci. 44, 671-678 (1997). https://doi.org/10. 1109/23.603731

[39] M. Anderson, et al., The Star time projection chamber: A Unique tool for studying high multiplicity events at RHIC.

Nucl. Instrum. Meth. A 499, 659-678 (2003). https:// doi.org/10.1016/S0168-9002(02)01964-2 [40] M. Beddo, et al. (STAR), The STAR barrel electro-magnetic calorimeter. Nucl. Instrum. Meth. A 499, 725-739 (2003). https://doi.org/10.1016/S0168-9002(02)

[41] L. Adamczyk, et al. (STAR), Measurements of Dielectron sNN = 200 GeV Production in Au+Au Collisions at from the STAR Experiment. Phys. Rev. C 92(2), 024912 (2015). https://doi.org/10.1103/PhysRevC.92.

[42] A. Adare, et al. (PHENIX), Detailed measurement of the e+e− sN N = pair continuum in p + p and Au+Au collisions at 200 GeV and implications for direct photon production. Phys.

Rev. C 81, 034911 (2010). https://doi.org/10.1103/

PhysRevC.81.034911

[43] B.I. Abelev, et al. (STAR), Systematic Measurements of Iden- tified Particle Spectra in pp, d+Au and Au+Au Collisions from STAR. Phys. Rev. C 79, 034909 (2009). https://doi. org/10.1103/PhysRevC.79.034909

Note: Figure translations are in progress. See original paper for figures.