Implementation of machine learning techniques to predict impact parameter and transverse spherocity in heavy-ion collisions at the LHC

Aditya Nath Mishra,[a,*] **Neelkamal Mallick ,**[b] **Sushanta Tripathy,**[c] **Suman Deb**[b] **and Raghunath Sahoo**[b,d]

[a]*Wigner Research Centre for Physics*
29-33 Konkoly-Thege Miklós str., 1121 Budapest, Hungary

[b]*Department of Physics, Indian Institute of Technology Indore*
Simrol, Indore 453552, India

[c]*INFN - sezione di Bologna*
via Irnerio 46, 40126 Bologna BO, Italy

[d]*CERN, CH 1211, Geneva 23, Switzerland*
E-mail: aditya.Nath.Mishra@wigner.hu, Neelkamal.Mallick@cern.ch,
Sushanta.Tripathy@cern.ch, Suman.Deb@cern.ch, Raghunath.Sahoo@cern.ch

Machine learning techniques have been quite popular recently in the high-energy physics community and have led to numerous developments in this field. In heavy-ion collisions, one of the crucial observables, the impact parameter, plays an important role in the final-state particle production. This being extremely small (i.e. of the order of a few fermi), it is almost impossible to measure impact parameter in experiments. In this work, we implement the ML-based regression technique via Gradient Boosting Decision Trees (GBDT) to obtain a prediction of impact parameter in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV using A Multi-Phase Transport (AMPT) model. After its successful implementation in small collision systems, transverse spherocity, an event shape observable, holds an opportunity to reveal more about the particle production in heavy-ion collisions as well. In the absence of any experimental exploration in this direction at the LHC yet, we suggest an ML-based regression method to estimate centrality-wise transverse spherocity distributions in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV by training the model with minimum bias collision data. Throughout this work, we have used a few final state observables as the input to the ML-model, which could be easily made available from collision data. Our method seems to work quite well as we see a good agreement between the simulated true values and the predicted values from the ML-model.

---

*Speaker

1. Introduction

The properties of hot and dense deconfined QCD matter, the Quark-Gluon Plasma (QGP), which is believed to be produced in high-energy heavy-ion collisions, are usually studied as a function of centrality classes of the collisions which are determined by the impact parameter (b) [1]. However, obtaining the impact parameter values from experiments is still challenging as its value ranges in few femtometers (fm). Thus, in experiments, the centrality classes are inferred from final state charged-particle multiplicities and sometimes from the transverse energy distribution. In the hindsight, it would benefit the experiments if one can successfully implement Machine Learning (ML) based technique to obtain the impact parameter in a precise way from the final state observables.

ML techniques have enabled the development of tools that have played an important role in the field of high-energy physics (HEP) along with in different fields of physics for decades [2, 3]. The impact parameter is one of the crucial physical quantities of heavy-ion collisions and it cannot be measured directly in experiments but might be inferred from observables at the final state. ML algorithms such as support vector machine (SVM) and neural network (NN) have shown a great success in learning the complexity of large datasets for determination of the impact parameter in heavy-ion collisions [4–9]. In the present work, ML-based regression technique via GBDT is used to obtain predictions for impact parameter and spherocity distributions in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 2.76 and 5.02 TeV using A Multi-Phase Transport (AMPT) [10]. For ML, we have used a python based ML package, named as scikit-learn software package [11]. We have specifically used the *GradientBoostingRegressor* module inside *sklearn.ensemble* framework.
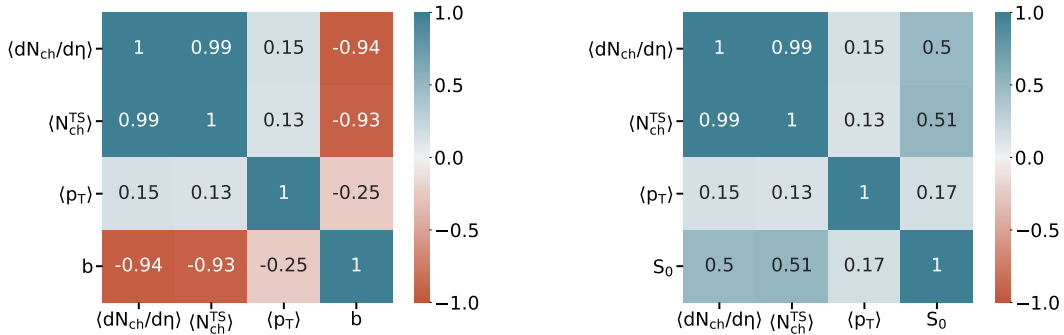

Figure 1: (Color Online) Correlation matrix of the input variables ($\langle dN_{\mathrm{ch}}/d\eta \rangle$, $\langle N_{\mathrm{ch}}^{\mathrm{TS}} \rangle$ and $\langle p_{\mathrm{T}} \rangle$) and target observables (impact parameter and transverse spherocity) in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV in AMPT model. The numbers show the correlation coefficients. The left panel shows the correlation matrix for impact parameter while the right panel shows the correlation matrix for transverse spherocity [12].

2. Results and Discussions

ML techniques are mainly used for classification, clustering and regression kind of problems. The problem addressed in this work is of supervised regression kind, *i.e.* for each set of the input variables, we have a finite numerical value as the target variable. Each set of the data refers to the final state observables of one event of the heavy-ion collisions. We have used charged-particle

multiplicity ($\langle dN_{ch}/d\eta \rangle$), charged-particle multiplicity in the transverse region ($\langle N_{ch}^{TS} \rangle$) and average transverse momentum ($\langle p_T \rangle$) as the input variables and the target variables as the impact parameter (b) and transverse spherocity (S_0). For the problem discussed here, a widely used gradient boosting decision trees (GBDT) algorithm has been chosen.

Figure 1 represents the correlation matrix for the input variables and the target variables for Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV minimum bias events. Here, left panel shows the correlation matrix for impact parameter while the right panel shows the correlation matrix for transverse spherocity. The numbers in the boxes represent the correlation coefficient which ranges from -1 to 1 and give the correlation strength between the intersecting variables in the matrix. The correlation coefficient (ρ) for two variables x and y is given by,

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \tag{1}$$

where cov(x, y) is the covariance and σ_x and σ_y are the standard deviations of x and y respectively.

From the values of ρ, from the correlation matrix of the input variables and impact parameter, it is evident that there is a significant anti-correlation between impact parameter and the $\langle dN_{ch}/d\eta \rangle$. Also, impact parameter is found to be (anti-)correlated with the $\langle p_T \rangle$ of an event. Figure 2(a) shows the predictions for impact parameter distribution using ML for Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV in AMPT model. The lower panel shows the ratio of predicted distribution to the true distribution. One can clearly see that the proposed ML framework with $\langle dN_{ch}/d\eta \rangle$, $\langle N_{ch}^{TS} \rangle$ and $\langle p_T \rangle$ as the input variables, does a nice job of predicting the impact parameter distribution in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV.
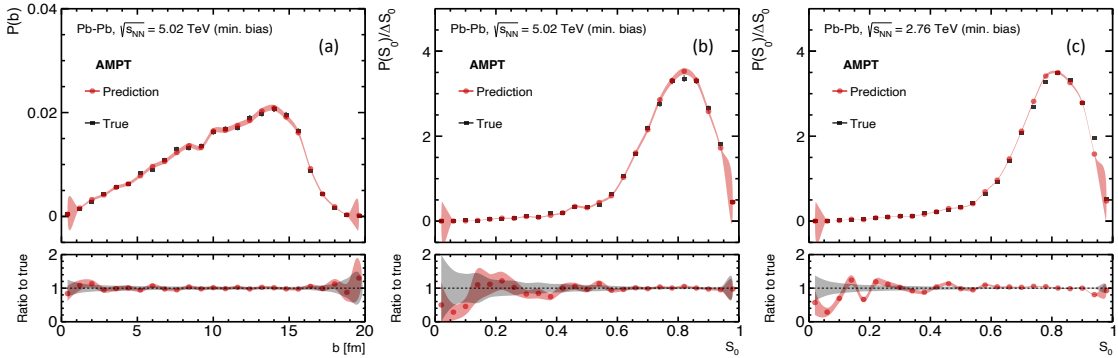

Figure 2: (Color Online) Predictions for the impact parameter for Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV and the transverse spherocity distributions for Pb-Pb collisions at $\sqrt{s_{NN}}$ = 2.76 and 5.02 TeV using ML-based regression technique via GBDT in AMPT model are shown. The quadratic sum of the statistical and systematic uncertainties are shown as a red-colored band for the predicted values. The statistical uncertainties in the true values are shown as bars. In the ratio, black-colored band denotes the statistical uncertainties in the true values while the red-colored band denotes the quadratic sum of statistical and systematic uncertainties [12].

The values of correlation coefficients (ρ), from the correlation matrix of the input variables and transverse spherocity in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV, suggest that there is a sufficiently high correlation between the input variables ($\langle dN_{ch}/d\eta \rangle$, $\langle N_{ch}^{TS} \rangle$ and $\langle p_T \rangle$) and the transverse spherocity of an event. Although, the correlation with $\langle p_T \rangle$ is small but it is still significant for a proper prediction

of transverse spherocity through ML-based regression technique via GBDT. Based on the mentioned inputs, predictions for transverse spherocity distribution in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV is shown in Fig. 2(b). Here. the predicted spherocity distribution is compared with the true ones obtained from AMPT. One can clearly see that the proposed ML-based regression technique via GBDT predicts the spherocity distribution accurately in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV. However, at low-spherocity regions, we see a deviation from the true distribution and this could be due to the fact that in heavy-ion collisions the statistics of having events with theback-to-back structure are expected to be quite less compared to events with isotropic in nature. Thus, we believe that this deviation could be due to limited statistics in the low spherocity region, which can also be seen by the statistical uncertainty black-colored band in the lower panel. In Fig. 2(c). We have also successfully predicted the spherocity distribution for Pb-Pb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV using the ML training from Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV in wide spherocity ranges. This suggests that the correlation of spherocity distributions with the input variables are quite similar across LHC energies. For more details about the methodology used in this article, one can see Ref. [12].

3. Summary

The impact parameter and transverse spherocity are two key observables in high-energy heavy-ion collisions. In this work, we implement the ML-based regression technique via Gradient Boosting Decision Trees to obtain a prediction of impact parameter and transverse spherocity in the midrapidity minimum bais Pb-Pb collisions at the LHC energies using A Multi-Phase Transport Model (AMPT) model. For this study, we use final state charged-particle multiplicity and mean transverse momentum as the input variables for the training of the ML algorithm.

In the absence of experimental measurements, we propose to implement ML-based regression technique to obtain transverse spherocity from the observed final state quantities in heavy-ion collisions. We would like to mention here that the ML-based training with the correlations of input observables using a MC model is quite useful when the MC model describes the input observables as close as possible to the experimental data. This method will be useful to handle the physics associated with unmeasured quantities in the experiment.

References

[1] S. A. Bass, M. Gyulassy, H. Stoecker and W. Greiner, J. Phys. G **25**, R1-R57 (1999).

[2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).

[3] A. Ortiz, A. Paz, J. D. Romo, S. Tripathy, E. A. Zepeda and I. Bautista, Phys. Rev. D **102**, 076014 (2020).

[4] S. A. Bass, A. Bischoff, C. Hartnack, J. A. Maruhn, J. Reinhardt, H. Stoecker and W. Greiner, J. Phys. G **20**, L21-L26 (1994).

[5] C. David, M. Freslier and J. Aichelin, Phys. Rev. C **51**, 1453 (1995).

[6] S. A. Bass, A. Bischoff, J. A. Maruhn, H. Stoecker and W. Greiner, Phys. Rev. C **53**, 2358 (1996).

[7] F. Haddad, K. Hagel, J. Li, N. Mdeiwayeh, J. B. Natowitz, R. Wada, B. Xiao, C. David, M. Freslier and J. Aichelin, Phys. Rev. C **55**, 1371 (1997).

[8] J. De Sanctis, M. Masotti, M. Bruno, M. D'Agostino, E. Geraci, G. Vannini and A. Bonasera, J. Phys. G **36**, 015101 (2009).

[9] F. Li, Y. Wang, H. Lü, P. Li, Q. Li and F. Liu, J. Phys. G **47**, 115104 (2020).

[10] Z. W. Lin, C. M. Ko, B. A. Li, B. Zhang and S. Pal, Phys. Rev. C **72**, 064901 (2005).

[11] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12 (2011).

[12] N. Mallick, S. Tripathy, A. N. Mishra, S. Deb and R. Sahoo, Phys. Rev. D **103**, (2021) 094031.