Decoding the nuclear symmetry energy event-by-event in heavy-ion collisions with machine learning

Yongjia Wang\textsuperscript{a}, Zepeng Gao\textsuperscript{a,b}, Hongliang Lü\textsuperscript{c}, Qingfeng Li\textsuperscript{a,d}\textsuperscript{*}

\textsuperscript{a}School of Science, Huzhou University, Huzhou 313000, China
\textsuperscript{b}Sino-French Institute of Nuclear Engineering and Technology, Sun Yat-sen University, Zhuhai 519082, China
\textsuperscript{c}HiSilicon Research Department, Huawei Technologies Co., Ltd., Shenzhen 518000, China
\textsuperscript{d}Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China

\textsuperscript{*}Corresponding author
Email address: lqf@zjhu.edu.cn (Qingfeng Li)

\textbf{Abstract}

Inferences of the nuclear symmetry energy from heavy-ion collisions are currently based on the comparison of measured observables and transport model simulations. Only the expectation values of observables over all considered events are used in these approaches, however, observables can be obtained event-by-event both in experiments and transport model simulations. By using the light gradient boosting machine (LightGBM), a modern machine-learning algorithm, we present a framework for inferring the density-dependent nuclear symmetry energy from observables in heavy-ion collisions on the event-by-event analysis. The ultrarelativistic quantum molecular dynamics (UrQMD) model simulations are used as training data. The symmetry energy slope parameter extracted with LightGBM event-by-event from test data also by UrQMD has an average spread of approximately 30 MeV from the truth, and is found to be robust against variations in model parameters. In addition, LightGBM can identify features that have the greatest effect on the physics of interest, thereby offering valuable insights. Our study suggests that the present framework can be a powerful tool and may offer a new paradigm to study the underlying physics in heavy-ion collisions.

\textbf{Keywords:} Nuclear symmetry energy, Heavy-ion collision, Machine learning, Event-by-event analysis

\textbf{Introduction.}—Nuclear symmetry energy, which quantifies the difference in binding energy per nucleon between isospin symmetric and asymmetric nuclear matter, is of crucial importance to the structure of nuclei (e.g., the neutron skin thickness and dipole polarizabilities), the dynamics of heavy-ion collision with radioactive beams (e.g., yield ratios between isospin partners), as well as properties of compact stars (e.g., neutron star tidal deformability and its mass-radius relationship)\textsuperscript{[13, 14, 15, 16, 17, 18]}. Deducing how the nuclear symmetry energy varies with density \(E_{\text{sym}}(\rho)\) from theories and experiments is one of the hot topics in nuclear physics and nuclear astrophysics. Heavy-ion collisions (HICs) with various species and energies offer a unique opportunity to access the nuclear symmetry energy at densities away from the saturation density \((\rho_0)\). However, \(E_{\text{sym}}(\rho)\) cannot be measured directly in experiment; usually it is deduced from the comparison between transport model simulations and experimental measurements\textsuperscript{[9, 10, 11, 12]}. Consequently, constraints on \(E_{\text{sym}}(\rho)\) from different models and experimental data diverge, especially above \(\rho_0\). There are a few reasons for this: (I) different assumptions and coding techniques are used in transport models—to disentangle these issues, the Transport Model Evaluation Project (TMEP) has been pursued since several years ago\textsuperscript{[13, 14, 15, 16, 17, 18]}; (II) different observables which may reflect properties of \(E_{\text{sym}}(\rho)\) at different densities are used to constrain \(E_{\text{sym}}(\rho)\)\textsuperscript{[20, 21, 22, 23]}; (III) the nuclear symmetry energy effects on observables in HICs can be easily concealed by large fluctuations caused by stochastic nucleon-nucleon collisions and by initial state fluctuations\textsuperscript{[24, 25, 26]}. To identify fingerprints of \(E_{\text{sym}}(\rho)\) on observables in HICs is one way to probe \(E_{\text{sym}}(\rho)\). Traditionally, averaged values of observables over all considered events are usually used to constrain \(E_{\text{sym}}(\rho)\) from the comparison of experimental measurements and transport model simulations. In this context, the extent of data used in traditional method is limited. Indeed, observables can be obtained event-by-event both in experiments and transport model simulations. Event-by-event data in HICs may consist of a huge amount of data of different observables for each event. These event-by-event data carries a lot of information about relevant physics and correlations among observables. Analyzing these event-by-event data and mining relevant information is a big challenge in traditional analysis. Can we build a framework that uses event-by-event data to infer \(E_{\text{sym}}(\rho)\) with modern data science techniques? Moreover, in traditional analysis, unsettled parameters (assumptions) of transport models have been found to affect the averaged values resulting in the divergence of the constrained \(E_{\text{sym}}(\rho)\)\textsuperscript{[27, 28, 29, 30, 21, 13, 31, 32]}. Therefore, it is desirable to infer \(E_{\text{sym}}(\rho)\) in a model-independent way.

Machine learning has been proven powerful for analyzing complex data in many branches of science\textsuperscript{[33, 34, 35, 36, 37, 38, 39]}. In this work, by using a modern machine-learning algorithm, we present a machine learning framework for inferring...
$E_{\text{sym}}(\rho)$ with event-by-event data, and the framework is found to be robust against variations in model parameters. This may open a new venue to study the underlying physics in HICs.

**Methodology.** Supervised learning is one of the machine learning techniques to map the relationship between input features and the output. Identifying $E_{\text{sym}}(\rho)$ from observables in HICs can be designed as a supervised learning task. One way to generate labeled data (observables that have been tagged with different $E_{\text{sym}}(\rho)$) is theoretical simulation. For this purpose, we use the ultrarelativistic quantum molecular dynamics (UrQMD) model. UrQMD has been widely employed for investigating HICs over a broad energy range \cite{40,41,42,43}. In presently used UrQMD model, the symmetry potential is derived from the Skyrme potential energy density functional \cite{44,45,46,47}, while the isoscalar component of the mean field potential inherits from the widely used soft and momentum dependent potential (SM, with the incompressibility $K_0=200$ MeV) in QMD-like models \cite{48,49}. Six different Skyrme interactions (Skz4, SLY230a, SV-sym32, SV-sym34, SkI2, and SkI1 \cite{50}) which yield very different $E_{\text{sym}}(\rho)$ are considered in the present work. $E_{\text{sym}}(\rho)$ and its slope parameter $L(\rho) = 3 p (\partial E_{\text{sym}}(\rho))/\partial \rho$ as a function of density are plotted in Figs. 1(a) and (b), respectively. It can be seen that $L(\rho)$ spans a wide range of values, fully covering the present uncertainty on $E_{\text{sym}}(\rho)$.

The SM nuclear potential, with the FP4 parameterization \cite{45} for the in-medium nucleon-nucleon cross section and an isospin-dependent minimum spanning tree (isoMST) method for cluster recognition, is applied as the default parameter set. In isoMST, two protons (two neutrons or neutron-proton pair) are considered to belong to the same fragment if their relative distance and momentum are smaller than $R_0^{pp} \ (R_0^{nn} \text{ or } R_0^{np})$ and $P_0$, respectively. The parameters adopted in the default parameter set are $R_0^{pp}=2.8$ fm, $R_0^{nn}=3.8$ fm, and $P_0=0.25$ GeV/c. In the present work, the training data consist of data generated with UrQMD using the default parameter set. Testdata1 is consistent with the training data. To validate the generalizability of trained machine learning algorithm, other parameter sets of UrQMD are also used to generate testing data, i.e., the free nucleon-nucleon cross section (Testdata2), larger parameters of $R_0^{pp}=3.5$ fm, $R_0^{nn}=4.5$ fm and $P_0=0.3$ GeV/c in isoMST (Testdata3), a hard and momentum dependent potential (HM, with the incompressibility $K_0=380$ MeV, Testdata4). We note here that, when varying model parameters in UrQMD, the magnitudes of observables are visibly influenced. For example, the yield of free protons in the case of Testdata1 is about 10% larger than that in Testdata3. The elliptic flow obtained with HM is about 50% stronger than that obtained with SM. When the nuclear symmetry energy is varied, the variations in observables are usually small. In practice, Au+Au collisions at beam energy of 0.4 GeV/nucleon with impact parameter $b=5$ fm are simulated; the number of events for generating training data and each testing data are 360000 and 50000 for each $E_{\text{sym}}(\rho)$, respectively. We have checked that the results are not changed with further increasing the number of events.

For data generation, 30 features (observables) related to the momenta ($p_1$, $p_2$, and $p_3$) of produced free protons and deuterons are used. They are ($p_1$), ($\langle p_1 \rangle$, $\langle p_2 \rangle$, $\langle p_3 \rangle$, $\langle v_1 \rangle$, $\langle v_2 \rangle$, $\langle v_3 \rangle$, $\langle v_1 \rangle$) and ($\langle v_1 \rangle$) of free protons and deuterons within rapidity windows $|y_0| < 0.5$ and $0.6 < |y_0| < 1.0$, and additionally ($|p_3|$) of free deuterons. Here, $p_i = \sqrt{p_{i1}^2 + p_{i2}^2}$, $v_1 = p_1/p_0$, $v_2 = (p_2^2 - p_1^2)/p_1^2$, and $v_3 = (4p_1^2 - 3p_2^2)/p_1^2$. The angle brackets indicate an average over all considered particles in each event. The normalized rapidity is defined as $y_0 = y/\gamma_{pro}$, with $\gamma_{pro}$ being the projectile rapidity in the center-of-mass frame. Observables with sub-index $p05$ (p61) and d05 (d61) referring to the results for free protons and deuterons within rapidity window $|y_0| < 0.5$ (0.6 < $|y_0| < 1.0$), respectively. To reveal how nuclear symmetry potential influence these observables, the distribution, scatter plot, and contour plot of ($p_1$)$_{p05}$, ($v_2$)$_{p05}$, ($v_1$)$_{d05}$, and

![Figure 1](image1.png) (Color online) The nuclear symmetry energy and its slope parameter are plotted as a function of density. Numbers listed in the right panel are the slope parameters at the saturation density.

![Figure 2](image2.png) (Color online) The diagonal plots show the distribution of each observable obtained with Skz4, SV-sym34, and SkI1. Upper off-diagonal scatter plots display the correlation between the horizontal and vertical labeled observables. Each scatter point denotes one single event; total 1000 events are displayed. Lower off-diagonal plots are the corresponding contour plots.
higher accuracy, faster training efficiency, as well as stronger interpretability because of LightGBM’s ability to handle large-scale data [51], thus it has been widely used and has achieved state-of-the-art performances in many tasks. Moreover, comparing to other machine learning algorithms, LightGBM yields a higher interpretability because of its decision-tree nature. In our previous works, the strong ability of LightGBM to refine nuclear mass models and determine impact parameter of HICs has been demonstrated [52][53][54]. All parameters in LightGBM are set to their default values; we have checked that the results are insensitive to parameters in LightGBM. To combat overfitting, the so-called L1 and L2 regularization [51] are implemented. In addition, testing datasets which constitute unseen data and data generated with different parameter sets in UrQMD are used to further prevent overfitting.

**Results**. - Both the training and testing data (i.e., Testdata1) generated with the same parameter set in UrQMD are tested, and the distribution of the predicted $L(p_0)$ is plotted in Fig. 2. First, similar to traditional analysis, if only the $p_1$ dependent $v_2$ observables are used, such as those listed in Fig. 1 (a), all of the predicted $L(p_0)$ approach to the global-averaged value [(5.8+44.3+57.0+81.2+106.4+159.0)/6=75.6 MeV] and the mean absolute error (MAE), which is the absolute difference between the true and the predicted $L(p_0)$, is as large as 38.1 MeV. Second, with 30 observables [see Fig. 2(b)], the distributions of predicted $L(p_0)$ are fairly separated and the MAE drops to 29.6 MeV. We note that this MAE value is larger than our previous result [53]. This is because the present study is event-by-event whereas our previous study treated the average over 100 events as one sample. We have checked, by using averaged observables over 100 events, that the MAE drops to 13.4 MeV with the present method. It is conceivable, with more variables included, that the widths of the distributions in Fig. 2(b) would narrow further and the peaks would be more cleanly separated. This demonstrates clearly the power of machine learning which can incorporate practically unlimited number of variables, which would be completely out of the question for a traditional analysis.

The mean value of the predicted $L(p_0)$, its standard deviation, and the MAE for Testdata1 using 30 variables are listed in Table I for the nuclear potentials we have studied. It is noteworthy for the cases of SLy230a, SV-sym32, SV-sym34, and SkI2 that the difference between the mean values of predicted $L(p_0)$ and the true values used in corresponding events is smaller than 24 MeV, while the differences in the cases of Sk4 and SkI1 (very soft and stiff ones) are as large as about 40 MeV. It is known that test dataset on extremes of the spectrum of the training datasets would yield event-by-event results significantly deviating from the truth. With a given real dataset, one would reiterate the procedure by including a wide spectrum of training data well embracing the test results. In reality, in the case of symmetry energy we study, it is known that $E_{\text{sym}}(\rho)$ is unlikely to be very soft or stiff [19][20][23][57][58], thus the iteration would quickly converge.

To illustrate the generalizability of the trained LightGBM model, we tabulate in Table II results for all testing datasets. Values of MAE from Testdata2, Testdata3, and Testdata4 are similar to that from Testdata1. This implies that the trained
LightGBM model is reliable even for inputs generated by UrQMD with different parameter sets, indicating small model-dependent systematic uncertainties. This suggests that the use of LightGBM to identify $E_{\text{sym}}(\rho)$ from more than a handful of observables has considerable potential.

Experimental data are often a mixture in impact parameter ($b$) which cannot be precisely determined. We thus test the trained LightGBM using data generated by UrQMD with two ranges of $b$: $0-2 \text{ fm}$ and $2-4 \text{ fm}$. The distributions of the predicted $L(\rho_0)$ are similar to those in Fig. 3, with the corresponding MAE of $32.7 \text{ MeV}$ and $31.2 \text{ MeV}$. These values are only slightly larger than that for fixed $b$ dataset, suggesting a strong potential of LightGBM in extracting $E_{\text{sym}}(\rho)$ from real experimental data. This is non-trivial as varying impact parameter can sometimes result in dramatic changes in observables.

Comparing to other deep learning algorithms, often treated as block boxes, LightGBM has an excellent degree of explainability because of its decision-tree-based nature. This is important as explainability often improves our knowledge about the relationship between the input features and the output. The Feature_importance (FI) technology of LightGBM and Shapley additive explanations (SHAP) are two popular feature attribution methods to identify the most important features that drive predictions. These two methods utilize different techniques to characterize the importance of features, although not equal, the identified important features are well correlated between the two methods. In our particular study, deuteron related features exhibit higher importance than protons in the determination of $L(\rho_0)$. This finding suggests that one can concentrate on deuteron-related observables to probe $E_{\text{sym}}(\rho)$. It can be understood from the fact that the distribution width of each deuteron-related observable is sensitive to $E_{\text{sym}}(\rho)$, as displayed in Fig. 4. LightGBM is robust against UrQMD model parameters in the nuclear potential, in-medium cross section, and cluster recognition. It is found that LightGBM works well with mixed impact parameters which ensures its applicability to real data analysis. The present framework can be generalized to studies of other physics in heavy-ion collisions, for example the extraction of the nuclear equation of state, the in-medium correction to the nucleon-nucleon cross section, the quark gluon plasma shear viscosity. It may open a new paradigm for studying the underlying physics in heavy-ion collisions.

We thank Fuqiang Wang for a careful reading of the manuscript and valuable communications. We acknowledge fruitful discussions with the TMEP group, M. B. Tsang, Yingxun Zhang, Kai Zhou, Jan Steinheimer, Yilun Du, Nan Su, Longgang Pang, Jun Su, Long Zhu, and Horst Stoecker. The authors are grateful to the C3S2 computing center in Huazhou University for calculation support. The work is supported in part by the National Natural Science Foundation of China (Nos. U2032145, 11875125, and 12147219), the National Key Research and Development Program of China under Grant No. 2020YFF0202002. The contributions of Hong-Liang Lu are non-Huawei achievements.

Table 1: The mean values of predicted $L(\rho_0)$ and their standard deviation $\sigma$ obtained with Gaussian fit. All units are in MeV.

| Table 1: Predicted $L(\rho_0)$ and standard deviation $\sigma$ obtained with Gaussian fit. All units are in MeV. |
|----------|--------|-------|--------|--------|-------|
| $L(\rho_0)$ | $L_{\text{true}}(\rho_0)$ | $\sigma$ | $L_{\text{true}}(\rho_0)$ | $\sigma$ | $L_{\text{true}}(\rho_0)$ | $\sigma$ |
| Sk24     | 5.8    | 44.1  | 16.8   | 43.3   | 16.1   | 38.4  | 16.4  | 48.0  | 17.0  |
| SLY230a  | 44.3   | 52.3  | 19.4   | 51.3   | 17.5   | 47.3  | 19.0  | 58.7  | 20.2  |
| SV-sym32 | 57.0   | 71.3  | 25.1   | 69.1   | 23.2   | 66.6  | 25.3  | 82.9  | 25.8  |
| SV-sym34 | 81.2   | 78.8  | 27.2   | 76.6   | 24.8   | 73.9  | 27.2  | 93.0  | 27.6  |
| Sk12     | 106.4  | 82.8  | 27.9   | 79.6   | 25.7   | 77.7  | 28.1  | 98.6  | 28.2  |
| Sk11     | 159.0  | 114.9 | 29.7   | 110.3  | 29.8   | 109.7 | 31.5  | 140.8 | 22.6  |

References

[1] B. A. Li, L. W. Chen and C. M. Ko, Phys. Rept. 464, 113-281 (2008).
[2] M. B. Tsang, J. R. Stone, F. Camera, P. Danielewicz, S. Gandolfi, K. Hebeler, C. J. Horowitz, J. Lee, W. G. Lynch and Z. Kohley, et al. Phys. Rev. C 86, 015803 (2012).
[3] H. Tong, P. Zhao, J. Meng, Phys. Rev. C 101, 035802 (2020).
[4] M. Oertel, M. Hempel, T. Klähn and S. Typel, Rev. Mod. Phys. 89, no.1, 015007 (2017).
[5] B. A. Li, P. G. Krastev, D. H. Wen and N. B. Zhang, Eur. Phys. J. A 55, no.7, 117 (2019).
[6] B. A. Li, B. J. Cai, W. J. Xie and N. B. Zhang, Universe 7, no.6, 182 (2021).
[7] P. G. Reinhard, X. Roca-Maza and W. Nazarewicz, Phys. Rev. Lett. 127, no.23, 232501 (2021).
[8] S. Huth, P. T. H. Pang, I. Tews, T. Dietrich, A. L. Fèvre, A. Schwenz, W. Trautmann, K. Agarwal, M. Bulla and M. W. Coughlin, et al. Nature 606, 276-280 (2022).
[9] Ono, A. Dynamics of clusters and fragments in heavy-ion collisions. Prog. Part. Nucl. Phys. 105, 139-179 (2019).
[10] J. Xu, Prog. Part. Nucl. Phys. 106, 312-359 (2019).
[11] M. Colonna, Prog. Part. Nucl. Phys. 113, 103775 (2020).
[12] C. W. Ma, H. L. Wei, X. Q. Liu, J. Su, H. Zheng, W. P. Lin, and Y. X. Zhang, Prog. Part. Nucl. Phys. 121, 103911 (2021).
[13] G. Jiang et al. [SrRIT and TMEP], Phys. Lett. B 813, 136016 (2021).
[14] J. Xu, L. W. Chen, M. B. Tsang, H. Wolter, Y. X. Zhang, J. Aichelin, M. Colonna, D. Cozma, P. Danielewicz and Z. Q. Feng, et al. Phys. Rev. C 93, 044609 (2016).
[15] Y. X. Zhang, Y. J. Wang, M. Colonna, P. Danielewicz, A. Ono, B. Tsang,
Figure 4: (Color online) The FI (upper) and SHAP (lower) values for each feature (variable) used in our study.