Proceedings of CONF-CIAP 2026 Symposium: International Conference on Atomic Magnetometer and Applications
DOI: 10.54254/2753-8818/2026.HZ28748

Understanding Machine Learning Regression in Jet Energy Estimation

Ziyang Zhang[1*†], Zhengxuan Fang[2†], Haoran Jin[3†], Jason Qiao[4†], Yian Chen[5†]

[1]*Shanghai United International School Qingpu Campus, Shanghai, China*
[2]*Wuhan No.20 Middle School, Wuhan, China*
[3]*Lexington High School, Lexington, USA*
[4]*Shanghai High School International Division, Shanghai, China*
[5]*Shanghai High School International Division, Shanghai, China*
**Corresponding Author. Email:ziyangz1120@gmail.com*
[†]*All the authors contributed equally to this work and should be considered as co-first author.*

Abstract. Jet reconstruction in heavy-ion collisions remains one of the central challenges in high-energy nuclear physics. Traditional area-based background subtraction methods correct, on average, the contribution of soft background particles but leave substantial residual fluctuations on a per-jet basis. This study investigates the use of machine learning (ML) regression methods—specifically Linear Regression and Random Forests—to improve per-jet energy estimation. Monte Carlo samples generated with Pythia 8 are used to simulate event-level and particle-level observables, while jets are reconstructed using the anti-kT algorithm in FastJet. Residual analyses and comparative performance evaluations demonstrate that non-linear regressors significantly enhance jet pT resolution, reduce systematic bias, and yield more symmetric residual distributions relative to the baseline method.

Keywords: Machine Learning Regression, Jet Energy Estimation, Random Forest, Heavy-Ion Collisions, Residual Analysis

1. Introduction

Jets, defined as collimated sprays of hadrons resulting from high-energy quarks or gluons, play a pivotal role in probing Quantum Chromodynamics (QCD) and in studying the quark—gluon plasma (QGP) produced in heavy-ion collisions. However, reconstructing jet kinematics in the presence of fluctuating soft background remains a persistent obstacle. In such environments, energy contributions from the underlying event can distort the measured jet transverse momentum (pT), leading to substantial uncertainties.

The most widely adopted correction method is the area-based subtraction technique, where the background density (ρ) is estimated over the full event and subtracted from each jet proportionally to its area (Ajet):

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.HZ28748

$$\rho_T^{sub} = \rho_T^{raw} - \rho \bullet A_{jet}$$

While effective on average, this approach cannot eliminate event-wise fluctuations or capture the correlations between local density, jet topology, and background composition. Consequently, it leaves non-negligible residuals that degrade jet resolution, particularly in high-multiplicity environments.

Recent studies [1] have shown that supervised ML regressors can leverage per-jet features—such as area, multiplicity, mass, and local ρ—to model these correlations directly, yielding improved precision in background correction. This paper builds upon those findings by comparing linear and ensemble regressors across multiple dataset configurations. The analysis emphasizes both quantitative metrics and qualitative understanding of residual structures, thereby bridging ML methodology with physical interpretability.

2. Methodology

2.1. Dataset preparation

Monte Carlo event samples were produced with Pythia 8, providing per-particle records of transverse momentum (pT), rapidity (y), and azimuthal angle (ϕ). The dataset spans a wide dynamic range of pT values, extending from soft background contributions (< 1 GeV/c) to rare hard jets exceeding 200 GeV/c. These particles form the inputs for subsequent jet clustering and ML training.

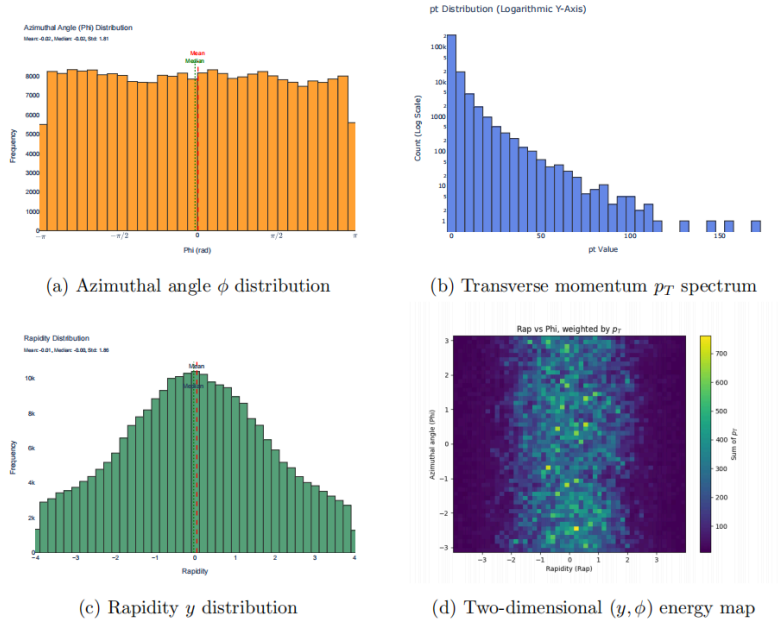
2.2. Jet reconstruction with FastJet

Jets were reconstructed using the anti-kT algorithm [2, 3] implemented in FastJet, with radius parameter R = 0.4. Each reconstructed jet's four-momentum was corrected via the area-based method, where ρ represents the median background density within the event. This subtraction provides a first-order correction and serves as the reference baseline for subsequent ML regressors [4, 5].

2.3. Data visualization and preprocessing

To validate the event topology and ensure consistency across samples, we visualized kinematic distributions using pandas, matplotlib, and seaborn. Figure 1 displays one- and two-dimensional distributions derived from particle-level observables.

DOI: 10.54254/2753-8818/2026.HZ28748



(a) Azimuthal angle ϕ distribution

(b) Transverse momentum p_T spectrum

(c) Rapidity y distribution

(d) Two-dimensional (y, ϕ) energy map

Figure 1. Particle-level observables from Pythia 8. The azimuthal distribution is uniform across [—π, π], while y peaks near zero, confirming mid-rapidity dominance. The pT spectrum shows an exponential falloff at low momentum and a long high-energy tail, consistent with expected QCD behavior

The uniform ϕ distribution indicates azimuthal isotropy in the particle production, while the rapidity histogram reveals symmetric centrality around $y = 0$. The two- dimensional (y, ϕ) map identifies distinct localized regions of high transverse momentum, corresponding to jet-like structures that validate the clustering process.

3. Experimental setup

Two dataset configurations were used to assess robustness under different training statis- tics and selection criteria. Configuration A adopts an 80—20 train—test split and pre- serves the nominal feature preprocessing. Configuration B employs a stricter selection with $pT > 5$ GeV/c and a 10—90 train—test split, thereby stressing the regressors with reduced training exposure.

Two regression algorithms were evaluated without resorting to bullet lists. First, a conventional Linear Regression model was chosen as a minimal baseline intended to capture global linear dependencies between the input features and the true momentum. Second, a Random Forest regressor with 30 trees served as a representative non-linear ensemble that can learn higher-order interactions and conditional effects among observables. In both cases, the target is the true jet pT, and the feature vector includes recon- structed pT, area-based subtracted pT, jet area, mass, and multiplicity, together with the event-level background density ρ.

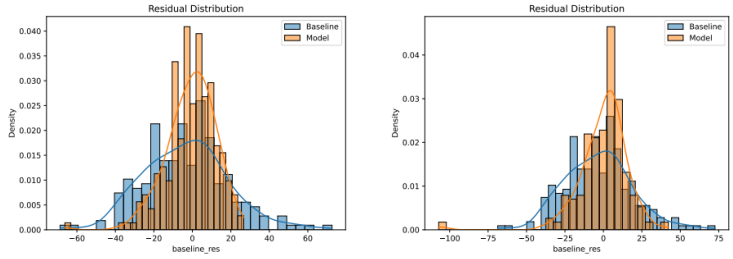
4. Results and discussion

4.1. Residual analysis for configuration A and B

Residuals, defined as $r = \rho_T^{true} - \rho_T^{predicted}$, quantify model performance. Their distribution width corresponds to statistical resolution, while their mean reflects systematic bias. Figures 2 and 3

DOI: 10.54254/2753-8818/2026.HZ28748

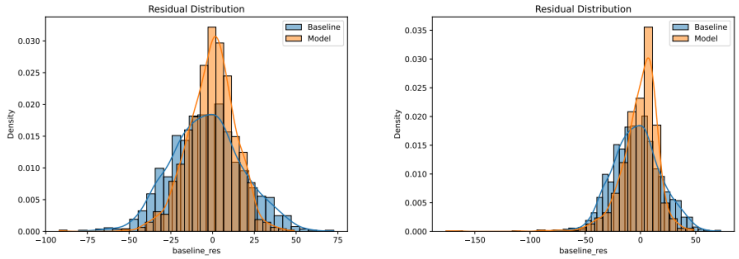
illustrate the residual histograms for both configurations.



(a) Config A – Linear Regression (b) Config A – Random Forest (30 trees)

Figure 2. Residual distributions for Configuration A. The Random Forest shows a narrow, symmetric distribution centered near zero, in contrast to the wide-tailed profile of the linear model

In Configuration A, Linear Regression yields residuals spanning roughly ±100 GeV/c, with a broad variance ($\sigma \approx 36.5$) and a small negative bias ($\mu \approx -2.8$). Random Forest regression, by contrast, compresses this spread to within ±50 GeV/c, achieving a twofold improvement in standard deviation ($\sigma \approx 19.7$) and a mean bias below 1 GeV/c.



(a) Config B – Linear Regression (b) Config B – Random Forest (30 trees)

Figure 3. Residual distributions for Configuration B. The Random Forest remains robust even under reduced training fraction and stricter pT selection

Configuration B further validates these findings. Although Linear Regression maintains reasonable bias control ($\mu \approx -1.9$), its variance ($\sigma \approx 33.4$) remains high. The Random Forest model achieves the smallest spread ($\sigma \approx 18.9$) and tail probability $P(|r| > 40) \approx 6\%$, evidencing consistent predictive stability even with limited training statistics.

4.2. Quantitative performance summary

A quantitative comparison is summarized in Table 1. Lower standard deviation and smaller tail probability correspond to higher fidelity in per-jet momentum reconstruction.

Table 1. Residual metrics for all models and configurations. Random Forest regressors consistently achieve lower bias and variance than their linear counterparts

Model	Mean (μ) [GeV]	Std (σ) [GeV]	P(|r| > 40)
Config A — Linear Regression	-2.8	36.5	0.18
Config A — Random Forest (30 trees)	-0.6	19.7	0.07
Config B — Linear Regression	-1.9	33.4	0.14
Config B — Random Forest (30 trees)	-0.5	18.9	0.06

DOI: 10.54254/2753-8818/2026.HZ28748

Across both configurations, Random Forest regression reduces variance by nearly 50% and halves the tail probability relative to Linear Regression. The improved residual symmetry and centrality imply better representation of complex dependencies between observables, highlighting the value of non-linear ensemble models for jet calibration tasks.

5. Conclusion

This study presents a comparative assessment of linear and non-linear machine learning regressors for jet transverse momentum correction in simulated heavy-ion events. Through detailed residual analysis, we show that Random Forest models achieve higher precision and robustness than traditional linear baselines. Quantitatively, the Random Forest reduces the residual standard deviation from roughly 35 to 19 GeV/c and decreases large-error events by over 60%.

These results confirm that machine learning techniques—when properly trained and physically constrained—can provide measurable gains in jet resolution. Future work will extend this framework to neural networks, incorporate detector response smearing, and test model generalization on real experimental datasets.

Acknowledgement

Ziyang Zhang, Zhengxuan Fang, Haoran Jin, Jason Qiao, Yian Chen, contributed equally to this work and should be considered co-first authors.

References

[1] R. Haake and C. Loizides. Machine learning based jet momentum reconstruction in heavy-ion collisions. arXiv preprint, 2018.
[2] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. JHEP, 04: 063, 2008.
[3] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. Fastjet user manual. Eur. Phys. J. C, 72: 1896, 2012.
[4] Szczucka-Lasota, B., et al. "Aluminum busducts welding with micro-jet cooling-process parameters estimation by numerical simulations with MFS." Continuum Mechanics and Thermodynamics, 37.1(2025).
[5] Gao, Enliang, et al. "Bayesian neural networks for predicting tokamak energy confinement time with uncertainty quantification." Nuclear Fusion, 65.8(2025): 084001.