Journal of Radioanalytical and Nuclear Chemistry (2025) 334:7231–7253
https://doi.org/10.1007/s10967-025-10422-6

Interpretable machine learning for rapid prediction and calibration of HPGe detector efficiency: a physics-informed approach and online platform

Cafer Mert Yesilkanat[1] · Necati Celik[2] · Ahmet Celik[3] · Ugur Cevik[4]

Received: 1 May 2025 / Accepted: 22 September 2025 / Published online: 8 October 2025

Abstract
This study introduces a novel, physics-informed, and calibration-friendly hybrid machine learning framework for the rapid and accurate prediction of the Full Energy Peak (FEP) efficiency in High-Purity Germanium (HPGe) detectors. To overcome the limitations of conventional "black-box" models, our two-stage approach first represents the FEP efficiency curve using a physically interpretable logarithmic polynomial. Subsequently, machine learning models were trained to predict the polynomial coefficients directly from the detector geometric parameters using a comprehensive dataset generated via Monte Carlo simulations. Among the various algorithms tested, the Generalized Linear Model yielded superior performance, achieving R^2 values of 0.975–0.992 for the coefficients. While raw model predictions showed expected variability, a key feature of our framework (a single-point calibration protocol using one known efficiency value) dramatically improved accuracy, reducing the mean absolute percentage error by an average of 80% on the independent test data. The calibrated model was further validated using an experimental detector. The entire framework was deployed as a user-friendly online platform, enabling researchers to instantly generate and calibrate efficiency curves. Our work successfully harmonizes interpretability, speed, and accuracy, offering a powerful tool for the design, optimization, and routine calibration of HPGe detectors.

Keywords HPGe detector · Machine learning · FEP efficiency · Monte Carlo simulation · GLM

Introduction

High-purity germanium (HPGe) detectors are widely used by laboratories to identify and quantify gamma-emitting isotopes in a wide variety of samples [1–3]. HPGe detectors are used in a wide range of applications, such as searching for dark matter [4], new neutrino properties [5, 6], unmanned aerial vehicles [7], and determining the activity of radioactive samples for nuclear diagnostics of Inertial Confinement Fusion (ICF) experiments [8–11]. The precise determination of the efficiency of the detection instrument under the specific measurement conditions of each sample is crucial for obtaining high-quality results. Given the diversity of conditions, such as measurement geometry, sample type, volume, and matrix, it is not possible to obtain complete calibration on an experimental basis alone [12]. With the increased availability of advanced algorithms, numerous software packages dedicated to performing efficiency calculations are available, such as Angle [13, 14], LabSOCS [15], ETNA [16], EFFTRAN [17], MEFFTRAN [18], GESPECOR [12], GEANT4 [19], EGS4 [20], EGSnrc [21]. Although the aforementioned validated Monte Carlo codes are available, they may not be accessible to everyone because their use requires a certain degree of training. In addition, depending on the complexity of the simulation geometry, the simulation times range from tens of minutes to hours. An alternative method to obtain the efficiency curve of a given HPGe detector relatively quickly would be extremely useful.

In recent years, with the widespread use of advanced algorithms, machine learning has permeated nearly every

✉ Cafer Mert Yesilkanat
cmyesilkanat@artvin.edu.tr

1 Science Teaching Department, Artvin Çoruh University, 08000 Artvin, Turkey

2 Faculty of Engineering and Natural Sciences, Gümüşhane University, 29100 Gümüşhane, Turkey

3 Department of Physics, Giresun University, 28200 Giresun, Turkey

4 Department of Physics, Karadeniz Technical University, 61080 Trabzon, Turkey

field of scientific research [22–30]. Given sufficient data, machine learning algorithms can learn input–output relationships and make predictions within certain error margins. Machine learning approaches that can assist or serve as alternatives to experimental or Monte Carlo simulations have begun to be explored in HPGe detectors. For example, Dey et al. [26] generated data using the FLUKA package program for a Broad-Energy HPGe (BEGe) detector and derived an empirical formula dependent on the detector geometry and energy, determining the coefficients of this formula through nonlinear regression. The same study also highlighted the effects of detector size ratios (e.g., diameter/height ratio) on the efficiency. Additionally, methods that directly derive formulas based on energy and distance [31] and black-box models based on artificial neural networks (ANNs) are present in the literature [32, 33]. In efficiency studies using ANNs for HPGe detectors, although the models often provide high accuracy, they have significant limitations in terms of interpretability and reliability assessment. This is because their internal workings are not transparent, and the relationships they establish with physical parameters are typically empirical and lack physical meaning [34, 35]. Therefore, the literature emphasizes the need for new methods that integrate physical knowledge and can quantify uncertainties to address the limited interpretability of ANNs [36, 37]. Moreover, when such direct prediction models contain systematic errors, it is often not possible for users to calibrate the model outputs, or separate corrections are required for each energy point [38].

To overcome these challenges, a paradigm shift from direct 'black-box' prediction to a more structured, physics-informed modeling framework is required. Such a framework should not attempt to learn the complex efficiency curve from scratch, but rather use machine learning to estimate the physically meaningful parameters of a well-established analytical model. By decomposing the problem in this way, two critical objectives can be achieved simultaneously: (1) the model's outputs (the coefficients) become interpretable in the context of detector physics, and (2) any systematic discrepancies between simulation and reality can be corrected through a simple calibration of these coefficients, rather than adjusting the entire curve point-by-point. This hybrid two-stage approach forms the core methodological innovation of our study.

In this study, we propose an interpretable, physics-based, and user-friendly hybrid model to address these limitations. Our model represents the full-energy peak efficiency curve of HPGe detectors using specific parameters and employs machine learning to estimate these parameters. Because the model outputs correspond to physically meaningful coefficients that describe the shape of the detector efficiency curve, the efficiency behavior of different detector geometries can be interpreted. Furthermore, a single known efficiency value at any point can be used to calibrate these coefficients, correct the systematic errors in the model, and simultaneously improve the predictions across the entire energy range. The proposed approach retains the speed and flexibility of the black-box model while incorporating the accuracy and reliability advantages of classical calibration methods. Finally, an online interface was developed to facilitate the easy use of the proposed model by researchers. Through this platform, users can input detector geometry parameters and any available experimental data to quickly and reliably obtain an efficiency curve for any energy value for HPGe detectors.

Materials and methods

Data production: Monte Carlo simulations

To train machine learning models, it is essential to create large and diverse datasets. In this study, the data were generated using MC simulations with the EGS 4 code to encompass the efficiency values for different energies and detector geometries, which would be extremely costly and difficult to obtain experimentally. For the EGS4 code, a MORTRAN code applied a cylindrical model, CCYSL [20]. This process includes the strategic placement of emission points inside the source. A RANLUX random number generator was used with the EGS4 code because it was shown to produce a relatively better distribution and longer sequence. For the model executed with EGS4, the efficiency data were divided into 10,010 energy bins, each with a width of 0.3 keV.

In the MC calculations, the calculation geometry was planned for an HPGe detector, and within the EGS4 framework, the detector volume was discretized into 919 cylindrical voxels (cells), each with material properties defined per reference geometry. The detector parameters used in these MC simulations are listed in Table 1. The calculated area was divided into 919 cells such that when this shape was rotated 360° around the z-axis (Fig. 1), a fully cylindrical geometry was obtained. Figure 1 shows the discretized geometry, with detector volume divided into voxels by a set of specific radial and axial divisions. The point source was modeled as a mono-directional, collimated photon beam. Every photon is emitted parallel to the central z-axis of the detector with no angular dispersion. [20, 21].

Using the EGS4 software package, we simulated the full-energy peak efficiencies for an HPGe detector across 20 gamma-ray energies ranging from 30 keV to 2.5 MeV. The simulation process began with a base detector model whose geometric parameters (listed in Table 1) precisely matched the specifications of a real HPGe detector. We then generated 100 variant geometries by randomly scaling these reference dimensions within physically realistic bounds (0.5 × to

Table 1 Reference parameters used for generating data for machine learning. Distances are in cm units; window material and relative efficiency were not used as variables; Detector type: n-type HPGe detector with 55% relative efficiency; and detector window is aluminum and kept constant in the current model.)

Detector Feature, abbreviation (unit)	Reference parameters
Detector-source distance, dsd (cm)	10
Window material	Al
Window thickness, wt (cm)	0.1
Window-crystal distance, wcd (cm)	0.4
Front dead layer thickness, wdt (cm)	0.07
Lateral dead layer thickness, ldt (cm)	0.07
Back dead layer thickness, bdt (cm)	0.0003
Crystal length, cl (cm)	4.360
Crystal radius, cr (cm)	3.135
Hole depth, hd (cm)	3.070
Hole radius, hr (cm)	0.515
Relative efficiency	55%

2 × scaling factors) while maintaining proper detector proportions. For each configuration, we performed independent MC simulations to calculate the efficiency values at all energy points.

The simulations incorporated all critical detector components, including the germanium crystal, dead layers, and entrance window, with experimentally validated geometries. We implemented point-source irradiation geometry. Each simulation tracked 10^8 photon histories to achieve statistical uncertainties of less than 0.5% in the calculated efficiency values. When combined with the inherent uncertainties in the interaction cross-sections, the total estimated uncertainty remained below 3%, making this dataset suitable for precise machine learning applications.

This approach offers several advantages: (1) the efficiency values at each gamma-ray energy are derived from first-principles radiation transport physics; (2) the varied geometries cover a realistic parameter space for different detector geometries; and (3) the statistical precision meets rigorous standards for model training. As shown in Fig. 1 and supported by the methodologies in references [26, 39], this simulation framework provides both physical accuracy and computational efficiency for generating the training data. The resulting dataset maintains direct correspondence with measurable detector properties while enabling comprehensive coverage of geometric variations that would be impractical to study experimentally.

Two-stage modeling framework

The efficiency values obtained from MC simulations can be processed either through direct machine learning mapping or via a physics-based model. In this study, we identified a critical limitation in approaches that rely solely on energy-dependent efficiency variations for a fixed detector geometry. Such methods risk learning artificial energy correlations rather than true physical relationships. These spurious correlations could significantly compromise both the generalizability and physical meaningfulness of the model. To overcome this limitation, we developed a novel two-stage modeling framework that shifts the learning objective from a direct efficiency prediction to the estimation of physically interpretable coefficients governing efficiency behavior. The proposed methodology first establishes an analytical representation of the efficiency curve using Eq. (1), a hybrid polynomial-inverse formulation specifically designed to capture the characteristic energy response of HPGe detectors. This physically grounded equation structure inherently preserves the expected

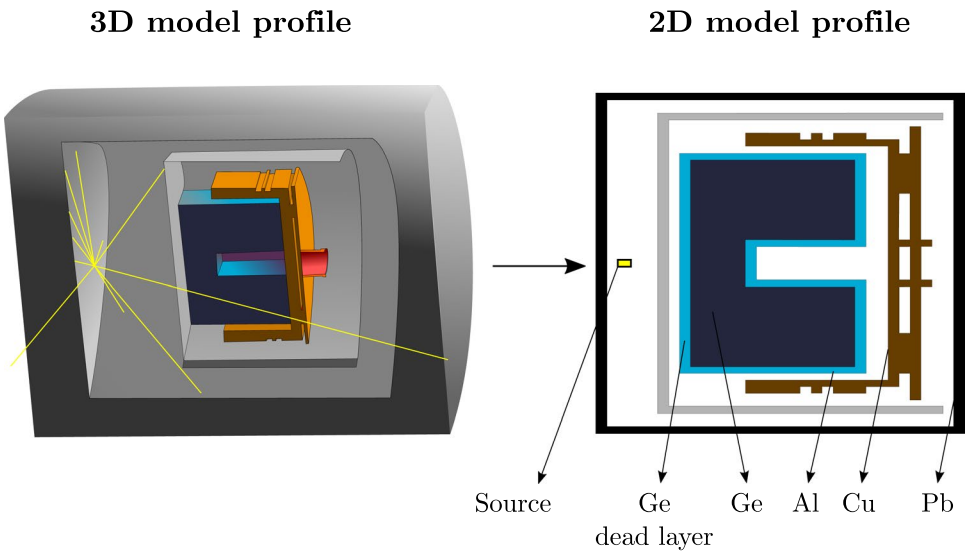


Fig. 1 Detector model used in MC simulation

asymptotic behaviors and energy dependencies observed in gamma-ray detection. In the second stage, machine learning algorithms were employed to predict the coefficients of the equation based on the detector geometry parameters. This approach provides several key benefits: it maintains a direct correspondence between the model parameters and measurable physical quantities, reduces the risk of unphysical predictions, and enables a meaningful interpretation of the results. The selected hybrid equation effectively represents the typical characteristics of the efficiency curve, including the rapid decrease at low energies and gradual roll-off at higher energies, while ensuring that all coefficients retain clear physical significance. By decoupling the physical representation from the parameter estimation process, our framework achieves superior generalizability compared with purely data-driven approaches. The physics-based equation constrains the solution space to physically plausible outcomes, whereas the machine learning component provides the flexibility to adapt to diverse detector geometries. This combination yields more reliable predictions, particularly for detector configurations outside the training distribution, because the model must adhere to fundamental physical principles rather than merely interpolating between data points.

$$\log[\varepsilon(E)] = a + b\log(E) + c\left[\log(E)\right]^2 + d\frac{1}{E} + e\frac{1}{E^2} \quad (1)$$

where $\varepsilon(E)$ is the energy-dependent efficiency, E is the photon energy, and a, b, c, d and e are model coefficients. The physical meanings of these coefficients are presented in Table 2.

Within the physics-informed modeling approach of this study, the machine learning process first predicted Table 2 coefficients based on the detector geometry parameters. These predicted coefficients were then substituted into Eq. (1) to calculate the efficiency values at any desired energy level.

Methodological process and performance metrics

In this study, a seven-step systematic process was implemented to predict HPGe detector full-energy peak (FEP) efficiency of HPGe detectors using a physics-informed two-stage machine learning approach. Each step is detailed below, following our updated modeling framework.

Step 1: Feature Definition and Engineering Methodology: Ten fundamental geometric parameters formed the initial inputs: detector-source distance (dsd), window thickness (wt), window-crystal distance (wcd), front dead layer thickness (fdt), lateral dead layer thickness (ldt), back dead layer thickness (bdt), crystal length (cl), crystal radius (cr), hole depth (hd), and hole radius (hr). Feature engineering techniques generated 17 additional derived variables with physical meaning, including crystal volume, surface area, volume-to-surface ratio, and length-to-diameter ratio, totaling 27 explanatory variables. Variable selection was based on correlation and multicollinearity analyses (complete definitions are provided in Supplementary Material 1).

Step 2: Monte Carlo-Based Data Generation: Using the reference GEM25P4-76 detector (Table 1), we created additional 100 geometric variants by randomly scaling each parameter (0.5–2.0 × range). For each geometry, the MC simulations generated efficiency data at 20 energy points (30–2500 keV), yielding 2020 observations covering most laboratory configurations.

Step 3: Extraction of Physical Coefficients: Each detector-energy dataset was fitted to a logarithmic efficiency polynomial via least-squares regression in R, deriving coefficients a-e. These coefficient vectors, which represent unique detector responses, are the target variables for machine learning prediction.

Step 4: Machine Learning Training Process: The dataset developed for the coefficient estimation was strategically divided into training (75%) and testing (25%) subsets to ensure robust model evaluation. We systematically applied

Table 2 The physical meaning of coefficients used in Eq. 2

Coefficients	Main physical resource(s)	Explanation	References
a	Effective solid-angle, effective volume, crystal–source placement	The energy-independent scaling factor of the detector geometry primarily determines the global efficiency	[40, 41]
b	Transition between Compton scattering + photoelectric apsorption	The negative slope in log(E), at medium energy level (≈0.1–1 MeV) describes the linear decrease in effi-cienfy in log scale	[40, 41]
c	Transition geometry between Photoelectric absorption → Compton scattering → Pair production	The term $\ln^2$(E), introduces concavity to the curve, physically capturing both the plateau width and the characteristic "shoulder" region	[40, 41]
d	Low-energy (< 80 keV) absorption; window and dead layer effect ($\mu \propto 1/E$)	The negative-signed 1/E term models the efficiency loss caused by thin windows/dead layers	[3, 31]
e	The sharp attenuation at very low energies (< 50 keV) arises from surface absorption effects ($\mu \propto 1/E^2$)	The $1/E^2$ term captures the sharpest energy dependence arising from the combined effects of dead layer, entrance window, and self-absorption	[3, 31]

eight advanced machine learning algorithms to the training data, including Generalized Linear Models (GLM), Random Forests (RF), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), Cubist models, Gaussian Process Regression (GPR), Bayesian Regularized Neural Networks (BRNN), and Multivariate Adaptive Regression Splines (MARS). Each algorithm's hyperparameters were meticulously optimized through an exhaustive tenfold cross-validation process combined with grid search techniques, with the specific objective of minimizing the mean squared error in the prediction of our target coefficients (a-e) [42]. This rigorous optimization procedure produced preliminary ("raw") models capable of accurately predicting the characteristic coefficients for any given detector geometry. Although these initial models successfully generated uncalibrated efficiency curves when their predicted coefficients were incorporated into Eq. (1), we observed a consistent scaling discrepancy between these theoretical predictions and our experimental/Monte Carlo reference data points. To address this systematic offset across the entire energy spectrum with maximum efficiency, we implemented an innovative single-point calibration protocol. This approach enabled the precise adaptation of our models to real detector characteristics while requiring only minimal experimental measurements; a single known efficiency value at any energy proved sufficient to properly scale all predictions throughout the energy range. The calibration process effectively eliminated the observed offset while maintaining the physical validity of our predictions, thereby demonstrating the practical utility of our physics-informed machine learning framework.

Step 5: Single-Point Calibration Protocol: The trained models were adapted to actual detector measurements using an efficient single-point calibration protocol. This practical correction method utilizes only one experimentally measured efficiency value at any energy point provided by the user. The calibration process calculates a dimensionless correction factor ($CF = \varepsilon_{actual}/\varepsilon_{predicted}$) that quantifies the ratio between the observed and predicted efficiencies. By applying this single scaling factor uniformly to all raw model outputs, the system simultaneously corrects both natural variations and systematic biases across the complete energy spectrum. This elegant approach yields detector-specific efficiency curves while maintaining several critical advantages: it preserves the physical relationships embedded in the original model, requires minimal experimental effort (only one calibration measurement), and automatically adjusts all the predicted values in a physically consistent manner. The resulting calibrated curves demonstrated excellent agreement with the reference measurements while retaining the computational efficiency of the machine learning framework, effectively bridging the gap between theoretical predictions and experimental reality.

While single-point calibration is sufficient to correct for the predominant scaling error, a multi-point calibration could be employed o enhance robustness in cases where the predicted shape of the efficiency curve exhibit minor but consistent deviations from the experimental data across the energy range, potentially due to unmodeled physical effects. This would involve a weighted least-squares adjustment of the predicted coefficients to fit multiple experimental points, futher reducing random uncertainties. The platform's architecture can be extended to support this advanced functionality for demanding applications where the highest possible accuracy is required.

Step 6: Computation of Performance Metrics: The predictive accuracy of the machine learning models was quantitatively assessed using four key statistical metrics: Mean Absolute Error (MAE, Eq. 2), and Root Mean Square Error (RMSE, Eq. 3), and the coefficient of determination (R^2, Eq. 4) [43], and Mean Absolute Percentage Error (MAPE, Eq.)). These metrics were specifically chosen to provide a comprehensive evaluation of the model performance from various perspectives. The MAE offers a robust measurement of the average prediction deviation, whereas the RMSE emphasizes larger errors through quadratic weighting. The concordance correlation coefficient evaluates both precision and accuracy relative to the line of perfect agreement, and the MAPE provides an intuitive percentage-based error interpretation. Each metric was calculated using standardized formulations to ensure a consistent evaluation across all detector configurations and energy ranges, enabling a direct comparison of the model performance while accounting for both systematic and random prediction errors. This multi-metric approach guarantees a thorough validation of the predictive capabilities of the models across different aspects of the efficiency curve estimation task.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{2}$$

$$\mathrm{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{4}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left(\frac{\left|y_i - \hat{y}_i\right|}{y_i}\right) \tag{5}$$

where n is the total number of points in cross-validation, y_i is the actual value, $\hat{y}_i$ is the estimated value, $\bar{y}$ is average of actual values.

Step 7- External Testing (Testing detector parameter sets) and Comparative Analysis: The most successful machine learning model was rigorously evaluated on ten completely independent test detectors that were excluded from all training phases. For each test case, we generated both pre- and post-calibration efficiency predictions and performed comprehensive comparisons with the reference Monte Carlo simulation results. To further validate the real-world applicability, we conducted identical testing procedures on an actual physical detector system. This two-tiered validation approach, using both simulated and experimental test cases, provides robust verification of the model's generalization capability beyond its training distribution. The comparison methodology maintained consistent evaluation metrics across all test scenarios, enabling a quantitative assessment of the effectiveness of the calibration protocol in bridging the gap between simulated predictions and real detector behavior. This stringent validation process confirmed the model's reliability for both theoretical studies and practical applications with physical detector systems.

Machine learning algorithm

Generalized linear model

Generalized Linear Models (GLMs) extend the classical linear regression theory by providing a flexible parametric framework capable of modeling response variables from any distribution within the exponential family [44]. At their core, GLMs establish a relationship between the expected value of the response variable ($E[Y|x]$) and a linear combination of explanatory variables through a carefully selected link function $g(.)$, as shown in Eq. (6).

$$g(E[Y|x]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{6}$$

This link function, chosen based on the response distribution (such as log for Poisson or logit for binomial distributions), enables the modeling of nonlinear relationships while maintaining a linear predictor structure [45]. The model parameters (β) are estimated through Maximum Likelihood Estimation (MLE), typically implemented computationally using the Iteratively Reweighted Least Squares (IRLS) algorithm for efficient convergence [46]. The inherent statistical regularity of exponential family distributions ensures that the resulting coefficient estimates are both consistent and asymptotically normal, allowing for the analytical derivation of standard errors and confidence intervals [47].

In the context of this study, the relationship between the detector geometry and polynomial coefficients was substantially linearized through comprehensive feature engineering. This transformed problem structure makes GLMs particularly advantageous compared to more complex tree-based alternatives, primarily because of their superior interpretability and enhanced generalization performance. The GLM implementation specifically utilized the Gaussian family with an identity link function for continuous coefficient prediction, supplemented by L2 regularization to address the potential multicollinearity among the derived features. Diagnostic analyses, including partial residual plots, confirmed the validity of the linearity assumptions in our modified feature space. This methodological choice reflects our dual emphasis on maintaining the physical interpretability of parameters while achieving computational efficiency, with GLMs demonstrating comparable predictive accuracy ($\pm 2\%$ MAE) to ensemble methods while providing transparent coefficient relationships that align with detector physics principles. The success of this approach underscores how proper feature engineering can make classical statistical methods competitive with modern machine learning techniques for well-structured physical problems.

Gaussian process regression

Gaussian Process Regression (GPR) represents a non-parametric, Bayesian approach that defines a probability distribution directly over a space of functions, eliminating the need for explicit functional form assumptions [48]. The fundamental premise of this model rests on the assumption that the target function values at any finite set of points follow a joint multivariate Gaussian distribution, expressed through a Gaussian Process (GP) prior, as shown in Eq. (7).

$$f(x) \sim \mathcal{GP}\left(m(x), k\left(x, x'\right)\right) \tag{7}$$

In this formulation, $m(x)$ denotes the mean function, and $k(x, x\prime)$ represents the covariance function (kernel) that fundamentally determines the model's expressive power by quantifying the input similarity and correlation. While the mean function is typically assumed to be zero in standard implementations, kernel selection (such as the radial basis function (RBF) kernel) provides a mechanism to incorporate prior knowledge about expected function characteristics, including smoothness properties [48]. The complete probabilistic framework emerges through the incorporation of Gaussian-distributed observational noise, enabling the analytical combination of the prior distribution and data likelihood via Bayes' theorem. This closed-form solution yields a posterior distribution that forms the basis of the predictions [49]. Unlike conventional regression techniques, GPR provides not only point estimates but also their associated predictive variance (or confidence intervals), reflecting the model's inherent uncertainty quantification capability, which is one of its most distinctive advantages. The variance of the predictive distribution naturally increases in regions

with sparse training data, offering valuable insights into prediction reliability that are particularly crucial in scientific applications.

In our detector efficiency modeling context, GPR's ability of GPR to capture complex, nonlinear relationships without overfitting while simultaneously providing uncertainty estimates makes it particularly valuable for coefficient prediction. The infinite differentiability assumption of the RBF kernel aligned well with our physically smooth efficiency curves, and the marginal likelihood optimization automatically balanced the model complexity with the goodness-of-fit. This Bayesian treatment also allows the natural incorporation of prior physical knowledge through kernel parameter constraints, creating an effective bridge between data-driven learning and domain expertise. The resulting predictions maintained physical plausibility, even for extrapolated detector geometries, demonstrating GPR's superior generalization capability of GPR compared with deterministic alternatives.

Multivariate adaptive regression spline

Multivariate Adaptive Regression Splines (MARS) is a flexible regression technique that captures global nonlinear behavior by systematically partitioning the dataset into local subspaces and constructing linear sub-models within each region [50]. The foundation of the method lies in its use of one-sided, piecewise-linear basis functions with hinge-like forms, which are mathematically expressed in Eq. (8) through two complementary formulations defined around the knot point (t). These basis functions adaptively respond to input variables through their distinctive segmented structures, enabling the modeling of complex nonlinear relationships without requiring predefined functional forms.

$$h(x_j, t) = \max\{0, x_j - t\} \text{ or } h(x_j, t) = \max\{0, t - x_j\} \quad (8)$$

The MARS algorithm, originally proposed by Friedman [51], operates through two fundamental phases. The forward selection phase begins by aggressively building an overfitted complex model through the sequential addition of variable-knot point pairs and their interactions, which most rapidly minimize the residual sum of squares [52]. This initial model construction is followed by a backward elimination phase that prunes the model using the Generalized Cross-Validation (GCV) criterion, effectively removing terms that either fail to enhance the generalization capability or introduce unnecessary complexity [53]. The resulting final model maintains strong interpretability by explicitly revealing local linear relationships within specific subspaces of the data space while simultaneously capturing nonlinear patterns without requiring complex functional representations [51].

In our study, MARS was specifically employed for polynomial coefficient prediction because of its unique ability to preserve the accurate prediction capabilities of linear models while adaptively modeling nonlinear components [54]. This dual capability is particularly valuable for addressing the mixed linear-nonlinear relationships between the detector geometry parameters and efficiency curve coefficients. The algorithm's automatic knot point determination successfully identified critical thresholds in the geometric parameter space, where efficiency characteristics transitioned between different regimes, while its piecewise linear formulation maintained the physical interpretability of the resulting sub-models. The implementation details included second-order interactions between the geometric parameters and energy-dependent terms, with the GCV penalty parameter optimized to balance the model complexity against the prediction accuracy across our diverse detector configurations.

Cubist model

The Cubist model represents an innovative rule-based hybrid methodology that effectively combines the partitioning logic of decision trees with localized linear regression models [55–57]. This algorithm operates by iteratively dividing the feature space into subregions through binary splits that maximally reduce the error variance, subsequently generating specialized linear regression models (Eq. (9)) for each terminal node (leaf). This unique architecture enables the representation of globally nonlinear relationships through interpretable piecewise linear surfaces [58], maintaining mathematical transparency while capturing complex patterns.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \quad (9)$$

Cubist distinguishes itself from conventional tree models through its two-stage refinement mechanism, which significantly enhances predictive performance. The first stage employs a boosting-like approach, where successive models (committees) are sequentially constructed to reduce the overall model variance [59]. The second stage incorporates a smoothing procedure that adjusts each prediction by blending it with the k-nearest neighbors (k-NN) from the training set that fall under the same rule, a process that softens abrupt transitions between rules while improving local prediction consistency [25, 57]. This hybrid architecture achieves an optimal balance between transparency and flexibility through several mechanisms [60]: (1) automatic variable selection and pruning capabilities that mitigate overfitting risks; (2) preservation of linear model interpretability within individual partitions; (3) incorporation of

ensemble methods' adaptability through committee models; and (4) local consistency enforcement via nearest-neighbor smoothing.

For our specific challenge of predicting polynomial coefficients from high-dimensional geometric features, Cubist was selected as an ideal middle ground between constrained models, such as GLM, and complex "black box" alternatives [58]. Its rule-based structure proved particularly effective in handling the piecewise-linear relationships observed between certain detector dimensions and efficiency curve parameters, whereas the committee-based approach provided the necessary flexibility to capture subtle nonlinearities. The model's automatic feature selection capability also helped prioritize the most physically meaningful geometric parameters from our engineered feature set, aligning with our emphasis on interpretable modeling. The implementation details included five committee members and seven nearest neighbors for smoothing, with parameters optimized through cross-validation to maintain physical plausibility while maximizing predictive accuracy.

Bayesian-regularized neural network

The Bayesian-regularized neural network (BRNN) represents a sophisticated integration of Bayesian statistical principles with artificial neural network architecture, fundamentally transforming the regularization process into an inherent component of the model rather than an external constraint [61, 62]. This approach establishes a probabilistic framework by assigning Gaussian prior distributions to network weights, thereby treating parameter estimation as a Bayesian inference problem rather than a purely optimization-driven process. Unlike conventional neural networks that primarily minimize the sum of squared errors (E_D), BRNN optimizes a composite objective function (Eq. (10)) that incorporates both the data fitting error and a weight magnitude penalty term (E_w), creating an intrinsic balance between model accuracy and complexity control.

$$E(w) = \beta E_D(w) + \alpha E_w(w) \tag{10}$$

The most significant innovation of BRNN lies in its automatic evidence-based optimization of the regularization parameters (α and β) through the Bayesian evidence framework [63]. This built-in mechanism fundamentally differs from traditional approaches by eliminating the need for external techniques, such as early stopping or manual hyperparameter tuning, while effectively managing both overfitting and error variance through posterior weight distributions. The implementation employs second-order optimization techniques, particularly the Levenberg–Marquardt algorithm, which enables the model to simultaneously learn the optimal network complexity and data noise characteristics directly from the training data [64]. This integrated approach bypasses the conventional requirements for cross-validation or separate validation sets while maintaining a robust generalization performance. Beyond its regularization advantages, BRNN's probabilistic foundation of BRNN provides several unique capabilities. The generation of posterior weight distributions enables meaningful uncertainty quantification for predictions [65, 66], offering confidence intervals that are particularly valuable in scientific applications. The model maintains the universal function approximation capabilities of standard neural networks while demonstrating superior performance on small- to medium-sized datasets, where conventional networks might overfit. However, this enhanced capability comes with computational trade-offs; the need to compute second-order Hessian matrices significantly increases the processing requirements for large network architectures, making careful design choices essential for practical implementation.

In our detector efficiency modeling context, BRNN was specifically selected to leverage the flexibility of neural networks while mitigating overfitting risks through its Bayesian framework. The automatic regularization adjustment proved particularly valuable given our moderately sized dataset of 100 detector configurations, where conventional neural approaches might struggle to balance complexity and generalizability. The resulting model provided not only accurate coefficient predictions but also valuable uncertainty estimates that informed our confidence in the efficiency curve projections across different geometric configurations. This combination of neural network flexibility with Bayesian rigor created an ideal solution for maintaining physical plausibility while capturing complex parameter relationships in the detector response characteristics.

Support vector regression

In recent years, the support vector machine (SVM) algorithm [67], distinguished by its superior performance in classification tasks, has been adapted to continuous-valued prediction problems under the designation of support vector regression (SVR) [68]. Being a method based on structural risk minimization optimization, which is an inductive principle, SVR is a supervised learning model with associated learning algorithms that are independent of data distribution [68]. Unlike linear regression, in SVR, estimation errors are maintained at a certain threshold level to reduce the negative effects of outlier data on the model. This threshold value (ε), which indicates the amount of up and down deviation from the separation curve, can be determined by solving the minimization problem shown in Eq. (11) according to the limits of Eq. (12) with the help of slack variables (ξ and ξ^*) and a kernel function [69, 70].

$$\text{Min}\left[\frac{1}{2}\omega^2 + C\sum_{i=1}^{m}(\xi + \xi^*)\right] \tag{11}$$

$$subject to\begin{cases} y_i - \omega\varphi(x_i) - b \le \varepsilon + \xi \\ \omega\varphi(x_i) + b - y_i \le \varepsilon + \xi^* \\ \xi, \xi^* \le 0, 1, 2, \cdots, m \end{cases} \tag{12}$$

where $\varphi(x)$ is nonlinear kernel function, x is an independent variable vector, ω and b are regulation parameters for the kernel function, and C is the penalty coefficient that determines the threshold range of the SVR model. The cited papers can be reviewed for the mathematical details of the SVR algorithm [68–72].

Random forest

Developed by Breiman [73], the Random Forest (RF) algorithm results from the combination of the bagging [74] and Random Subspace [75] methods to increase the prediction performance of decision trees. In the RF approach, there are many randomly generated decision trees rather than only one. Thus, each tree created affects the RF results at specific weight levels. Therefore, the RF is an ensemble learning technique [39].

In the RF application, the entire dataset is randomly divided into two main parts; two-thirds serve as training data (in-bag) for the learning process, and one-third serves as validation data (out-of-bag) to determine the level of learning [76]. Subsequently, using the bootstrapping approach, many random decision trees are generated [77]. The formation of each tree is determined by randomly selecting input variables at the nodes from the root to the leaves. The output value obtained by the decision trees was compared with the actual results, and a weight was assigned to each tree considering the error rates. Thus, while the average of all trees is considered for the final prediction, the trees that provide the correct prediction are more heavily weighted, and the effects of trees that show an incorrect prediction on the result are reduced. In addition, it can be determined that the input variables causing the correct branching of the trees, which yield an accurate prediction, will also have a high level of significance.

The RF algorithm incorporates substantial randomness and prevents the occurrence of overfitting because both the variables and trees are composed of different randomly selected sub-datasets [78–80].

Extreme gradient boosting

The eXtreme Gradient Boosting (XGBoost) algorithm is a scalable and optimized version of the gradient boosting machine (GBM) algorithm created by Friedman [81]. The XGBoost algorithm, introduced by Chen and Guestrin [82], represents a significant advancement over traditional gradient boosting methods, primarily distinguished by three key innovations [83]. First, it employs a closed-form optimization approach that incorporates second-order derivatives (Hessian information) of the objective function when determining the leaf weights and other parameters. This results in faster and more stable convergence compared to conventional gradient-based methods that rely solely on first-order derivatives [39]. Second, XGBoost integrates an L_2-based structural regularization term that penalizes both the tree depth and leaf count, directly embedding model complexity control within the objective function. This eliminates the need for manual pruning while automatically optimizing the bias-variance trade-off. Third, engineering optimizations, including column subsampling, a parallelized split-finding algorithm, adaptive default direction handling, and out-of-core memory management for missing values, enable XGBoost to process millions of observations within seconds, making it exceptionally scalable [82].

In this study, XGBoost was selected to address the challenge of predicting polynomial coefficients because of its ability to efficiently capture complex, potentially nonlinear interactions within our high-dimensional derived feature space. The algorithm provided superior flexibility compared to Generalized Linear Models (GLMs) while offering finer-grained control over model tuning than Random Forests (RFs). Its built-in regularization mechanisms effectively mitigate overfitting risks, ensuring robust performance across diverse detector geometries. These attributes make XGBoost particularly suited for our problem, where interpretable but adaptable modeling is essential for bridging geometric parameters to efficiency curve characteristics. The implementation leveraged hyperparameter optimization (e.g., learning rate and tree depth) to balance physical plausibility with predictive accuracy, ultimately achieving MAE values below 2% for coefficient estimation. This performance underscores the utility of XGBoost in physics-informed machine learning applications that require both computational efficiency and nuanced handling of high-dimensional feature interactions.

Software sources Machine learning analyses and calculations in this study were performed in the R programming environment [84]. The R library files used in this study were kernlab for the SVR algorithm [85], ranger for the RF algorithm [86], xgboost for the XGBoost algorithm [87], brnn for the BRNN algorithm [88], cubist for the Cubist algorithm [89], glmnet for the GLM algorithm [90], earth for the MARS algorithm [91], mgpr for the GPR algorithm [92], and shiny [93] for the online computational front-end tool. Furthermore, the caret [94, 95] and dplyr [96] library files

were employed for data processing, encompassing preparation, separation, and optimization for analysis.

All analyses were performed using the latest R version as of June 2025, on a system with a 36-core Intel® Core™ i9-14900HX processor, an RTX4090 graphics card, and 64 GB of RAM, with parallel processing support.

Results and discussions

Statistical profile of geometric and derived parameters

A comprehensive statistical analysis of the fundamental and derived geometric parameters revealed critical insights into the physical processes governing the HPGe detector efficiency curves. As shown in Table 3, the distributions of measurable parameters, such as source–detector distance (mean dsd = 11.837 cm), window thickness (mean wt = 0.089 cm), and window–crystal distance (mean wcd = 0.594 cm), cluster around typical laboratory configurations. All geometric variants were generated by applying one random scale between 0.5 × and 2.0 × uniformly to every dimension of the baseline detector, thereby preserving aspect ratios. By design, more sampled cases lie above the baseline than below (the range above 1.0 is wider), which reflects the chosen bounds rather than statistical skewness, while still ensuring the inclusion of extreme but physically meaningful scenarios (see min–max in Table 3). The inclusion of these boundary cases is particularly valuable for

Table 3 Statistical summary of geometric/derived parameters and model coefficients

Variable	Formula	Mean	SD	Min	Max	Skewness	Kurtosis
Source-Detector Distance	dsd	11.837	4.589	5.048	19.781	0.322	1.822
Window Thickness	wt	0.089	0.043	0.026	0.195	0.866	2.858
Window-Crystal Distance	wcd	0.594	0.198	0.204	0.975	0.166	2.094
Front Dead Thickness	fdt	0.0050	0.0043	0.0006	0.0139	0.6473	1.9465
Lateral Dead Thickness	ldt	0.039	0.043	0.001	0.137	0.640	2.019
Back Dead Thickness	bdt	0.0045	0.0050	0.0000	0.0139	0.5397	1.7555
Crystal Length	cl	4.752	2.126	1.505	9.512	0.560	2.370
Crystal Radius	cr	4.440	1.461	1.801	7.534	0.186	2.004
Hole Depth	hd	1.869	1.753	0.181	5.711	0.690	2.057
Hole Radius	hr	0.785	0.573	0.141	1.997	0.757	2.088
Crystal Volume	$\mathrm{cl} \times \pi \times \mathrm{cr}^2$	320.95	250.55	27.86	1431.05	1.59	6.39
Radius/Length Ratio	cr / cl	1.162	0.710	0.307	3.939	1.203	4.341
Solid Angle Ratio	$\mathrm{cr}^2/(\mathrm{dsd}^2+\varepsilon)$	0.259	0.277	0.011	1.463	1.773	6.382
Front Surface Area	$\pi \times \mathrm{cr}^2$	68.58	42.52	10.19	178.31	0.63	2.35
Total Side Surface	$2 \times \pi \times \mathrm{cr} \times \mathrm{cl}$	131.76	72.81	23.25	399.79	0.98	4.00
Total Surface Area	$2 \times \pi \times \mathrm{cr}^2 + 2 \times \pi \times \mathrm{cr} \times \mathrm{cl}$	268.92	137.97	59.35	721.81	0.66	3.07
Volume/Surface Ratio	$(\mathrm{cl} \times \pi \times \mathrm{cr}^2)/(2 \times \pi \times \mathrm{cr}^2 + 2 \times \pi \times \mathrm{cr} \times \mathrm{cl})$	1.060	0.310	0.470	1.983	0.470	2.827
Geometric Efficiency	$\mathrm{cr}^2/(4 \times (\mathrm{dsd} + \mathrm{cl})^2)$	0.028	0.028	0.002	0.142	1.829	6.515
Dead Volume Ratio	$(\pi \times \mathrm{hr}^2 \times \mathrm{hd})/(\mathrm{cl} \times \pi \times \mathrm{cr}^2)$	0.031	0.052	0.000	0.242	2.233	7.555
Effective Crystal Volume	$\mathrm{cl} \times \pi \times \mathrm{cr}^2 - \pi \times \mathrm{hr}^2 \times \mathrm{hd}$	311.04	246.03	27.79	1423.16	1.65	6.66
Total Dead Layer	fdt + ldt + bdt	0.049	0.036	0.011	0.138	0.852	2.451
Window/Length Ratio	wt / cl	0.021	0.012	0.006	0.073	2.038	8.376
Window Effect	wt /(dsd + wt + wcd)	0.008	0.005	0.002	0.026	1.229	4.646
Crystal/Hole Radius Ratio	cr /(hr + ε)	9.358	7.735	1.424	35.439	1.794	6.043
Length/Depth Ratio	cl /(hd + ε)	6.189	5.560	0.771	28.306	1.273	4.422
Effective Radius	cr − hr	3.655	1.458	0.833	6.542	0.147	2.018
Effective Length	cl − hd	2.883	1.574	-0.988	6.315	-0.017	2.103
coefficient *a*		6.026	5.003	-3.883	18.251	0.395	2.723
coefficient *b*		-2.280	1.357	-5.686	0.301	-0.481	2.739
coefficient *c*		0.092	0.081	-0.056	0.302	0.548	2.835
coefficient *d*		-127.96	73.88	-290.29	19.59	-0.15	2.26
coefficient *e*		720.12	1287.21	-1815.73	2902.94	-0.18	1.74

modeling low-energy photon absorption (< 100 keV), where the entrance window and gap (wcd) regions disproportionately affect the baseline and slope of the efficiency curve through energy-dependent attenuation. Although small in absolute magnitude (microns to tens of microns), the dead-layer thicknesses (fdt, ldt, bdt) vary enough to materially impact low-energy efficiency because they act as primary absorbers before photons reach the active crystal volume.

The distribution characteristics of both the fundamental (dsd, wt, wcd, fdt, ldt, bdt, cl, cr, hd, hr) and derived parameters (volume, surface area, ratio, effective dimension, dead volume fraction, etc.) presented in Table 3 quantitatively illustrate the diversity of the physical processes that shape the HPGe detector efficiency curves. These parameters were systematically generated using EGS4-based Monte Carlo (MC) simulations by scaling the reference geometry values, resulting in a comprehensive geometric parameter space that is difficult to achieve experimentally. While the mean values of directly measurable parameters, such as source-detector distance (dsd = 11.837 cm), window thickness (wt = 0.089 cm), and window-crystal distance (wcd = 0.594 cm), cluster near typical laboratory conditions, their right-tailed (bounded by the 0.5 × –2 × scaling; see min–max in Table 3) distributions and high kurtosis values confirm the inclusion of extreme scenarios in the dataset. This ensures that the model accounts for significant variations in the low-energy photon absorption effects, particularly in the window and gap regions, which critically influence the baseline level and slope of the efficiency curve in the low-energy range [97].

Despite their small absolute magnitudes, dead layer thicknesses (e.g., fdt ≈ 0.005 cm, ldt ≈ 0.039 cm) have a substantial impact on efficiency, as they act as the first barriers to low-energy photons before they reach the active crystal volume [3]. The distribution patterns in Table 3 indicate that these thicknesses vary significantly across detector parameter sets, leading to distinct breaks in the efficiency curves, particularly in the low-energy region. Macro-geometric parameters, such as crystal length (cl) and radius (cr), primarily influence the total active volume and geometric acceptance angle, thereby affecting the intercept of the efficiency curve in log space [98, 99]. While their mean values typically fall within 4–5 cm and 3–5 cm, respectively, their extended distributions (up to ~ 9.5 cm and ~ 7.5 cm) explain the significant variations observed in the polynomial coefficients (a, b, c) across different detector parameter sets.

The central hole dimensions (hd and hr) reduce the effective crystal volume, predominantly controlling the slope and concavity of the efficiency curve in the mid-to-high energy range [99, 100]. The distribution of the dead volume ratio (mean ≈ 0.031, SD ≈ 0.052) is right-tailed (consistent with the bounded 0.5 × –2 × scaling space; see Table 3) and confirms that the hole effects can reach unusual levels in some detector parameter sets, leading to a wide spread in the $1/E$ and $1/E^2$ terms (coefficients d and e). This is numerically reflected in the high standard deviation of the coefficient e relative to its mean (SD/Mean ≈ 1.8).

The derived parameters, including crystal volume ($cl \times \pi \times cr^2$), surface areas, volume-to-surface ratio, geometric efficiency $\left(cr^2/\left[4 \times (dsd + cl)^2\right]\right)$, and effective dimensions, exhibited right-tailed leptokurtic distributions (kurtosis > 3; bounds reported in Table 3), indicating the inclusion of rare but physically influential configurations. For instance, crystal volumes exceeding 1430 cm^3 shift the log-efficiency intercept upward, thereby altering the overall curve level [99, 101]. Similarly, the geometric efficiency and solid-angle-like ratios, despite their small mean values, exhibit extreme right-tailed distributions that play a key role in shaping the contours of the efficiency curve in the mid-energy range, where Compton scattering dominates [102, 103].

Statistical profile of model coefficients

Figure 2 presents the detailed probability distributions of the polynomial coefficients (a, b, c, d and e) that define the detector efficiency curves in log-energy space. When interpreted alongside the descriptive statistics in Table 3, these distributions provide a comprehensive understanding of the structural characteristics governing the efficiency curve behavior.

The first three coefficients (a, b, c) exhibited relatively narrow distributions with low skewness and kurtosis values. Coefficient b, with a mean of -2.28 and standard deviation of 1.36, shows a near-symmetric distribution (skewness = − 0.48), consistent with the expected monotonic decrease in efficiency with increasing $log(E)$. Similarly, the coefficient c demonstrates a small positive mean (≈ 0.09) with limited variation (SD ≈ 0.08), indicating a consistent concave curvature across the detector configurations. Coefficient a, representing the log-scale intercept, averaged 6.03 with moderate variability (percent coefficient of variation; CV = 83%). Together, these coefficients maintain the fundamental "skeleton" of the efficiency curves across low-to-medium energies, showing relative independence from the detector geometry.

In contrast, coefficients d and e, which govern the high-energy behavior, exhibited significantly broader distributions. Coefficient d (mean = -127.96, SD = 73.88, CV > 58%) and especially coefficient e (mean = 720.12, SD = 1287.21, CV ≈ 180%) exhibit heavy-tailed distributions (kurtosis = 1.74) indicating extreme sensitivity to geometric variations. This variability correlates strongly with high-CV geometric parameters, such as effective crystal volume (SD = 246.03) and dead volume ratio (CV > 160%),

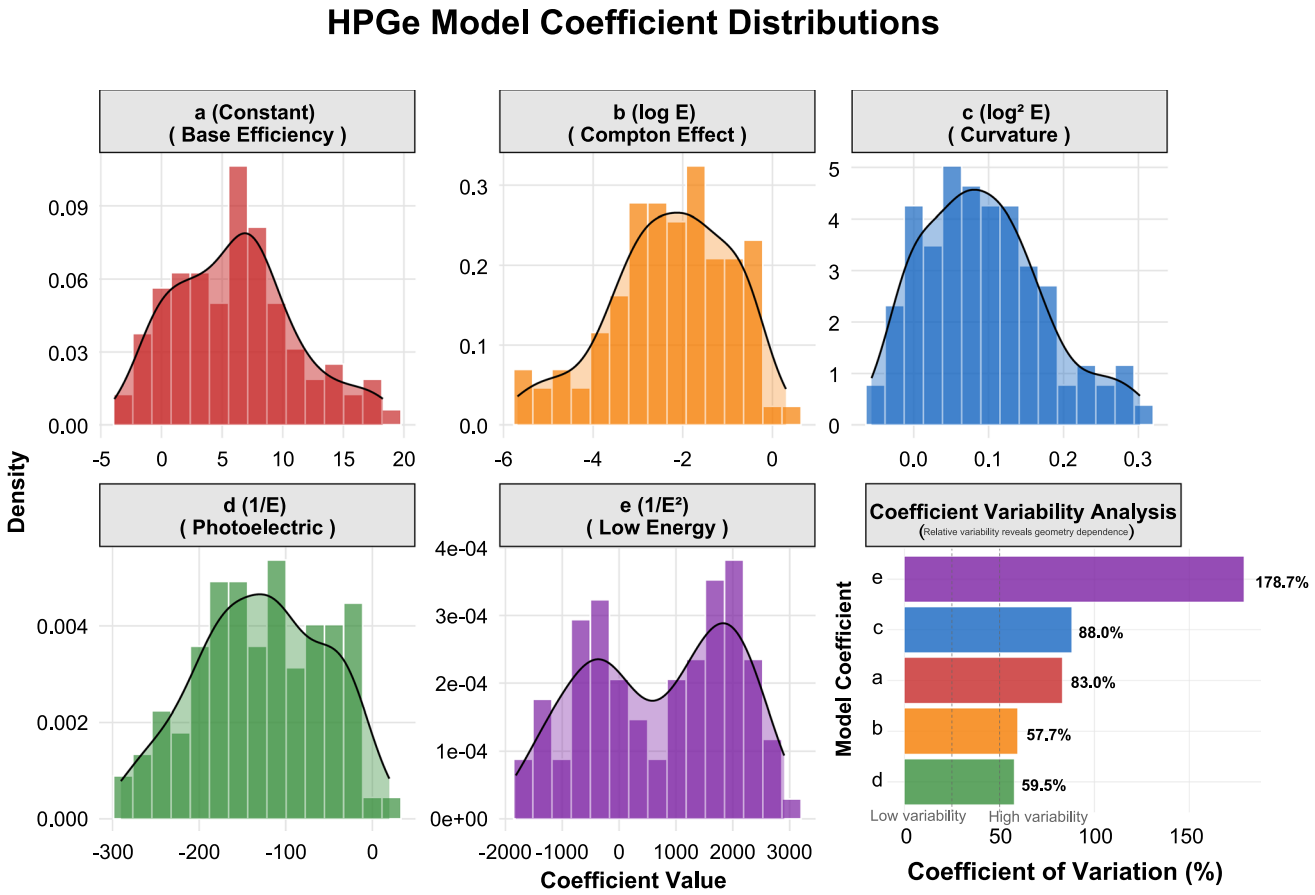


Fig. 2 Distribution of model coefficients and analysis of the coefficient of variation

which predominantly affect the sharp efficiency drop at high energies.

These findings reveal a fundamental dichotomy: the low-to-medium energy range demonstrates stable and predictable behavior. The high-energy segment shows extreme sensitivity to microscopic geometric variations, providing critical guidance for future detector calibration strategies and suggesting that high-energy calibration points may offer greater leverage for constraining the entire efficiency curve through physical scaling factors or offsets. The results quantitatively validate the hybrid modeling approach, in which machine learning predicts these physics-based coefficients rather than raw efficiency values.

ML process: training and testing

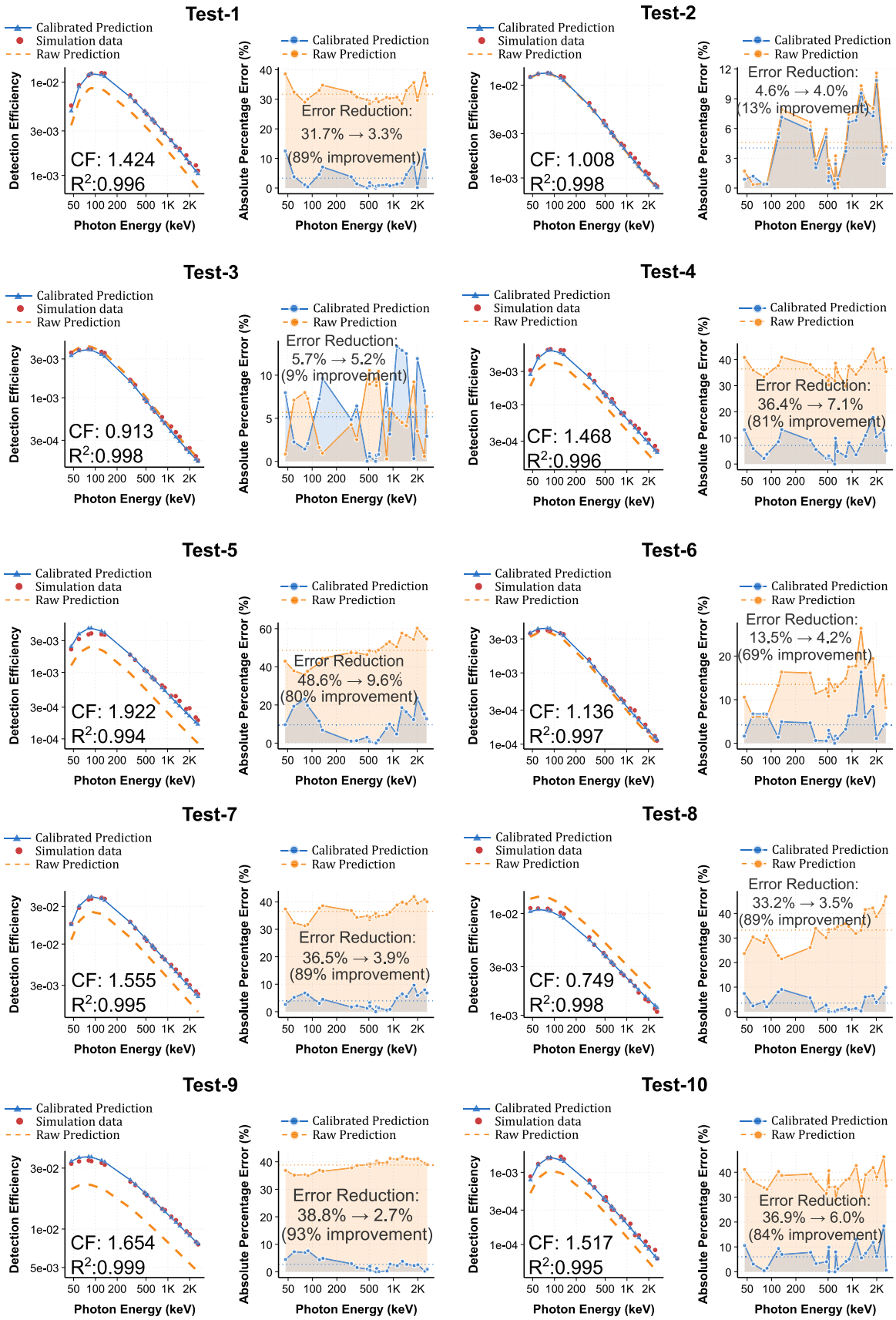
The machine learning modeling process was conducted to predict the five polynomial coefficients (a, b, c, d, e) defining the detector efficiency curve using 27 fundamental and derived geometric variables obtained through Monte Carlo simulations, the statistical summary of which is presented in Table 3. Each coefficient was treated as a separate target variable, and training was performed using eight different regression algorithms, as detailed in Sect. "Methodological process and performance metrics". The dataset was divided into training (80%) and test (20%) sets, with the training process employing 10-repeat, tenfold cross-validation to enhance model stability and generalization capability. The reported training performance metrics represent averages across 100 validation runs, whereas the test performance was evaluated on an independent holdout dataset. The performance metrics for the training and test results of each coefficient are presented in Table 4.

The analysis of Table 4 reveals that the GLM model generally demonstrated superior performance in terms of coefficient prediction accuracy. The GLM model achieved the highest or near-highest R^2 values for all coefficients, with particularly notable results. For coefficient a, the GLM model achieved a test R^2 of 0.987, demonstrating its ability to produce low-variance and highly linear predictions. Although the SVR and Cubist models also performed well with R^2 values of 0.961 and 0.984, respectively, the GLM's test RMSE (0.614) and MAE (0.483) were significantly lower. Similarly, for coefficient b, the GLM achieved a test R^2 of 0.983 with an RMSE of 0.172 and an MAE of 0.136, outperforming the other algorithms. The MARS

Table 4 Training and test performance metrics for the coefficients in the equation that reflect the typical behavior of the detector efficiency curve of ML algorithms. (* the best performance model)

Coefficients	Models	Train RMSE	Train R^2	Test RMSE	Test R^2	Test MAE
a	*GLM	1.054	0.965	0.614	0.987	0.483
	RF	1.357	0.948	1.842	0.857	1.461
	SVR	0.965	0.969	0.931	0.961	0.593
	XGBoost	1.404	0.922	1.605	0.890	1.207
	Cubist	0.932	0.968	0.767	0.984	0.609
	BRNN	1.144	0.955	0.794	0.979	0.602
	GPR	1.056	0.961	0.853	0.976	0.651
	MARS	1.088	0.958	0.809	0.973	0.621
b	*GLM	0.278	0.964	0.172	0.983	0.136
	RF	0.328	0.959	0.389	0.908	0.318
	SVR	0.266	0.968	0.256	0.957	0.145
	XGBoost	0.348	0.944	0.378	0.912	0.301
	Cubist	0.267	0.968	0.257	0.968	0.206
	BRNN	0.333	0.951	0.231	0.975	0.179
	GPR	0.289	0.957	0.243	0.972	0.179
	MARS	0.308	0.953	0.246	0.979	0.188
c	*GLM	0.020	0.954	0.013	0.975	0.010
	RF	0.021	0.954	0.025	0.893	0.021
	SVR	0.020	0.949	0.022	0.914	0.014
	XGBoost	0.022	0.928	0.027	0.862	0.020
	Cubist	0.018	0.954	0.016	0.967	0.011
	BRNN	0.021	0.941	0.013	0.969	0.010
	GPR	0.021	0.946	0.017	0.961	0.013
	MARS	0.021	0.943	0.017	0.972	0.015
d	*GLM	13.374	0.971	7.178	0.991	5.754
	RF	16.812	0.965	23.069	0.910	18.287
	SVR	12.338	0.977	11.654	0.978	7.348
	XGBoost	19.710	0.940	19.872	0.923	14.924
	Cubist	11.421	0.981	10.891	0.985	7.535
	BRNN	14.249	0.967	14.297	0.966	12.164
	GPR	13.650	0.969	10.455	0.985	8.211
	MARS	14.478	0.967	10.415	0.981	7.207
e	*GLM	190.036	0.981	124.103	0.992	107.430
	RF	300.376	0.951	408.254	0.923	312.664
	SVR	159.094	0.986	139.682	0.991	110.669
	XGBoost	265.758	0.959	431.694	0.901	322.692
	Cubist	162.215	0.987	129.046	0.994	99.996
	BRNN	171.712	0.983	177.079	0.987	136.356
	GPR	198.176	0.980	164.897	0.988	135.287
	MARS	149.785	0.987	140.721	0.991	124.428

model was the closest, with an R^2 of 0.979, but had higher RMSE (0.246) and MAE (0.188). For coefficient *c*, the GLM maintained excellent performance with test $R^2 = 0.975$, RMSE = 0.013, and MAE = 0.010. Although Cubist and BRNN showed comparable results, GLM's linear interpretability of GLM made it preferable. The performance of coefficients *d* and *e*, which describe the behavior in the high-energy region, is particularly critical. For coefficient *d*, GLM achieved $R^2 = 0.991$ with RMSE = 7.178 and MAE = 5.754, which is clearly superior to the other models. For coefficient *e*, GLM's $R^2 = 0.992$ with RMSE = 124.103 and MAE = 107.430 demonstrated consistent dominance, despite competitive R^2 values from Cubist and SVR models. These results confirm GLM's optimal balance of accuracy and interpretability of GLM for this physics-based modeling task, where feature engineering successfully linearized the


Test-1
Calibrated Prediction
Simulation data
Raw Prediction
Detection Efficiency
CF: 1.424
R²:0.996
Photon Energy (keV)
Absolute Percentage Error (%)
Error Reduction:
31.7% → 3.3%
(89% improvement)
Test-2
CF: 1.008
R²:0.998
Error Reduction:
4.6% → 4.0%
(13% improvement)
Test-3
CF: 0.913
R²:0.998
Error Reduction:
5.7% → 5.2%
(9% improvement)
Test-4
CF: 1.468
R²:0.996
Error Reduction:
36.4% → 7.1%
(81% improvement)
Test-5
CF: 1.922
R²:0.994
Error Reduction
48.6% → 9.6%
(80% improvement)
Test-6
CF: 1.136
R²:0.997
Error Reduction:
13.5% → 4.2%
(69% improvement)
Test-7
CF: 1.555
R²:0.995
Error Reduction:
36.5% → 3.9%
(89% improvement)
Test-8
CF: 0.749
R²:0.998
Error Reduction:
33.2% → 3.5%
(89% improvement)
Test-9
CF: 1.654
R²:0.999
Error Reduction:
38.8% → 2.7%
(93% improvement)
Test-10
CF: 1.517
R²:0.995
Error Reduction:
36.9% → 6.0%
(84% improvement)

◂**Fig. 3** Comparison of raw and calibrated model predictions for different detector geometries using simulation efficiency data. (The calibrated predictions were generated by using a single, mid-energy data point (661.6 keV) from each detector's respective Monte Carlo simulation as the reference calibration value.)

relationships between geometric parameters and efficiency curve characteristics. The model's consistent outperformance of more complex algorithms highlights the value of parametric regression when physical relationships are appropriately structured.

These results demonstrate that the relationship between the detector geometry and the coefficients can be largely characterized by a linear approximation, whereas more complex models (such as RF, XGBoost, and BRNN) exhibit inconsistencies in test performance or risks of overfitting as they attempt to capture nonlinear relationships. The simplicity and physical interpretability of the GLM provide a distinct advantage in terms of model transparency, which is critically important in detector physics studies.

The primary reason for GLM's superior predictive performance of GLM for each coefficient is believed to stem from its compatibility with the detector efficiency curve structure defined in Eq. (1). In this study, the GLM was selected for predicting the coefficients because it yielded better results than the other models when explaining the detector efficiency curves. Within this framework, GLM-based prediction models are considered the optimal choice for optimizing detector design and the physical calibration processes.

The model's consistent outperformance of more sophisticated algorithms confirms that proper feature engineering successfully linearized the underlying physical relationships between the geometric parameters and efficiency characteristics. This approach maintains rigorous connections to first principles while benefiting from computational efficiency, which is a crucial combination for practical applications in radiation detection systems. The demonstrated performance ($R^2 > 0.97$ for all coefficients) validates the GLM as both physically meaningful and statistically robust for this specific modeling challenge.

Experimental validation of model predictions with independent data

As established in the previous section, the GLM was identified as the optimal method for predicting the polynomial coefficients of detector efficiency curves from geometric parameters. The raw efficiency values generated using these GLM-predicted coefficients through the logarithmic polynomial equation (Eq. 1), defined in the first section of this article, was calculated for each desired energy value. To verify the practical performance of the developed model, a comprehensive two-stage validation process was conducted using completely unseen data that the model had not encountered during training. This validation approach systematically evaluated the predictive capability of the model by first testing it against new Monte Carlo simulation data, followed by experimental measurements from physical detector parameter sets. The process specifically examined both raw predictions and calibrated outputs, with particular attention to the model's ability to maintain accuracy across different detector configurations and energy ranges while preserving the physical interpretability inherent in the GLM framework.

The initial validation phase was conducted using 10 completely new test detector parameter sets generated from Monte Carlo simulations that the model had never encountered previously. The predictions were evaluated in two distinct scenarios: (1) raw predictions obtained directly from the model based solely on the geometric parameters of each detector, and (2) calibrated predictions optimized through single-point calibration. The results of these tests are presented graphically in Fig. 3, with the detailed numerical performance metrics listed in Table 5.

The comparative diagrams in Fig. 3 show the simulation data (ground truth), raw predictions, and calibrated predictions for each of the detector parameter sets. The analysis of these results revealed significant deviations between the raw predictions and actual simulation values for certain detector parameter sets, with notable examples including Test-1 (MAPE: 31.74%), Test-4 (MAPE: 36.41%), Test-5 (MAPE: 48.64%), and Test-9 (MAPE: 38.84%). These discrepancies stem from the inherent uncertainties in the coefficient predictions based solely on the geometric parameters.

Table 5 quantitatively demonstrates the variable and often substantial error rates in raw predictions across different detector parameter sets, with Test-5 showing particularly high errors (i.e., MAPE: 48.64%). However, the implementation of the proposed single-point calibration method using actual efficiency measurements at one energy point dramatically reduced these errors. For Test-5, calibration decreased the MAPE to 9.60%, representing an 80.25% improvement in the accuracy. Similarly, Test-9 showed remarkable enhancement, with the MAPE decreasing from 38.84 to 2.73% (92.97% improvement). Across all test detector parameter sets, calibration achieved an approximate 70% average error reduction, demonstrating near-perfect agreement between the calibrated predictions and simulation values.

Furthermore, we conducted a detailed residual analysis to scrutinize the performance of the single-point calibration protocol and to validate our framework's ability to reproduce the fundamental shape of the efficiency curve. While metrics like MAPE quantify the average error, a residual plot reveals whether any systematic, energy-dependent deviations remain after calibration. Figure 4 presents the percentage residuals for the ten independent test detectors. The analysis of these residuals yields two critical insights. First, for the vast

Table 5 Error metrics for raw and calibrated predictions of test detector parameter sets and post-calibration improvement rates

Detector	Calibration factor	MAPE raw	MAPE calibrated	Improvement MAPE	R^2
Test-1	1.424	31.74	3.33	89.49	0.996
Test-2	1.008	4.62	4.03	12.67	0.998
Test-3	0.913	5.66	5.18	8.57	0.998
Test-4	1.468	36.41	7.09	80.53	0.996
Test-5	1.922	48.64	9.60	80.25	0.994
Test-6	1.136	13.48	4.23	68.62	0.997
Test-7	1.555	36.50	3.94	89.20	0.995
Test-8	0.749	33.19	3.52	89.40	0.998
Test-9	1.654	38.84	2.73	92.97	0.999
Test-10	1.517	36.88	6.03	83.65	0.995

majority of detector configurations and energy points, the residuals are distributed within the ± 10% band, with many falling within ± 5%. This quantitatively confirms the high accuracy of the calibrated model for practical applications and demonstrates the profound effectiveness of the single-point calibration in eliminating the primary scaling error of the raw, geometry-based predictions. Second, the residual diagnostics provide an objective assessment of the model's remaining limitations. In some cases (e.g., Test1—Test4) in the Fig. 4, a mild systematic bias is evident as most residuals fall slightly below the zero line, indicating a small overestimation that persists even after calibration. This suggests the presence of small, second-order discrepancies in the predicted curve shape, likely arising from complex physical effects not fully captured by the logarithmic polynomial. However, the random scatter of residuals around the zero-line in most cases (e.g., Test-5, Test-8, Test-9) confirms that the model is not afflicted by a major, systematic shape-prediction failure. Ultimately, this residual analysis provides compelling evidence that our physics-informed approach does significantly more than a simple rescaling; it preserves the physical integrity of the efficiency curve's shape with high fidelity. The presence of minor, interpretable residual trends in challenging cases does not detract from the model's overall utility. Instead, it underscores the model's transparency and provides a clear pathway for future enhancements, such as the implementation of a multi-point calibration protocol to correct these finer shape deviations for applications demanding the highest tier of precision.

The second validation stage focused on comparing the model predictions with the experimental data obtained from an actual detector system, and the results are presented in Fig. 5. Comparative analysis revealed significant discrepancies between the raw model predictions and measured values (MAPE: 19.7%), whereas calibrated predictions following single-point correction demonstrated remarkable alignment with the actual detector performance (MAPE: 4.6%). This outcome confirmed the reliability and accuracy of the model under real experimental conditions. The graphical results in Fig. 5 illustrate how effectively the calibration process reduces the prediction errors for the physical detector data, mirroring the success observed in simulation-based tests. The post-calibration efficiency curve achieved exceptional

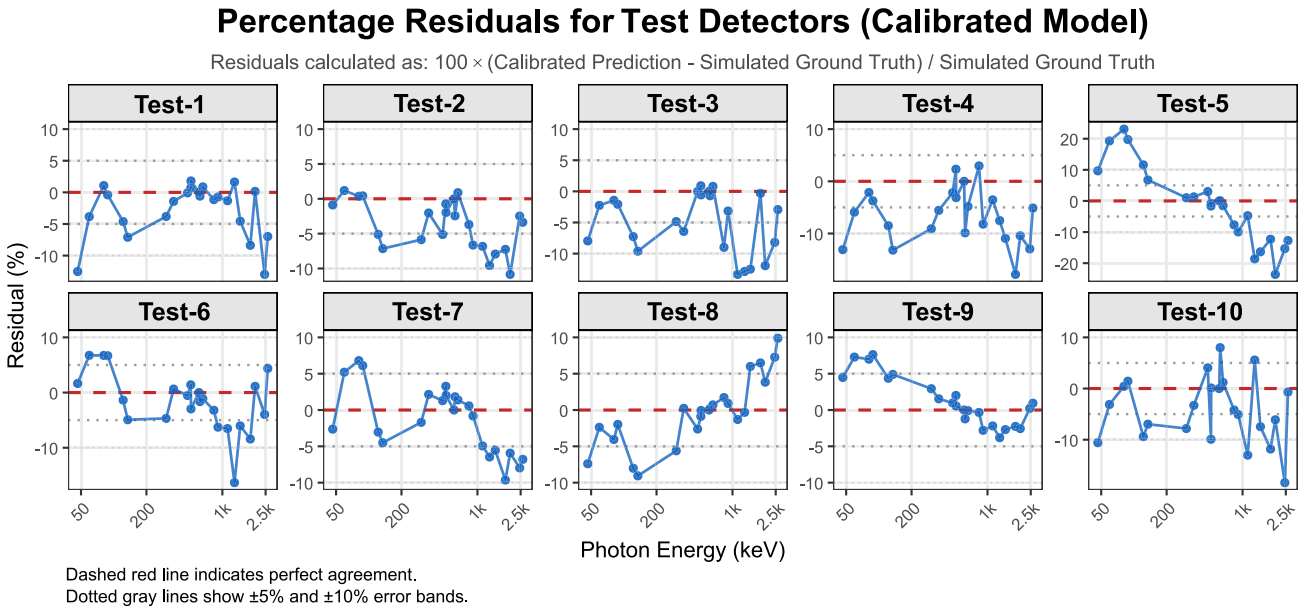


Fig. 4 Residual plots of the single-point calibrated prediction values for test detectors

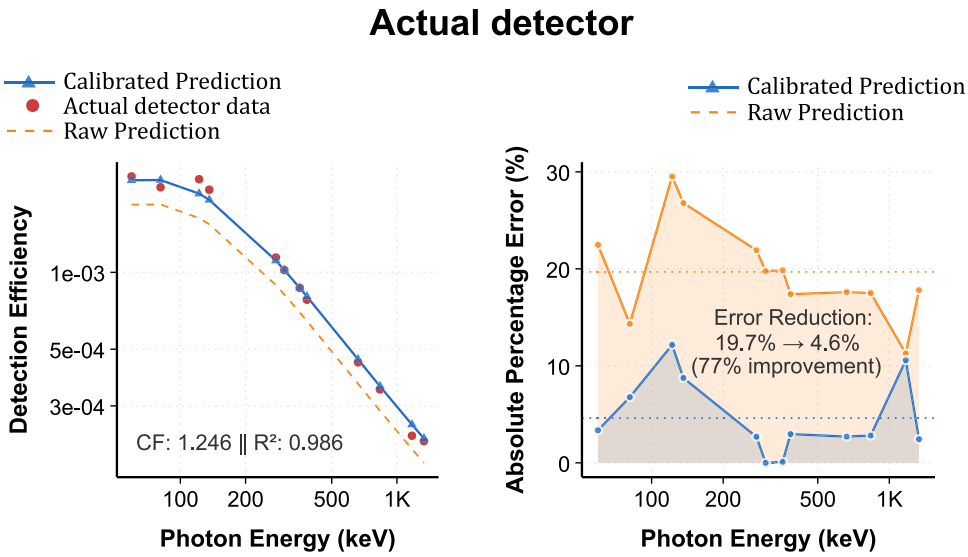


Fig. 5 Comparison of simulation efficiency data versus raw and calibrated model predictions for different detector geometries

agreement with the experimentally measured values, demonstrating near-perfect consistency. These findings validate that the geometry-based predictions of the model can be successfully implemented in practical applications, maintaining both theoretical rigor and measurement accuracy. The effectiveness of the calibration protocol in bridging the gap between simulated predictions and experimental results is particularly valuable for real-world detector characterization and optimization processes.

This two-stage validation process conclusively demonstrated the reliability and practical utility of the developed GLM-based model. The results highlight the effectiveness of single-point calibration in mitigating the inherent variations and systematic biases present in raw model predictions, thereby significantly enhancing the practical value of the model for real-world applications. The validation confirmed that the proposed methodology serves as an effective tool for both simulation-based analyses and actual detector design/calibration processes. These findings establish that the model functions as a reliable and efficient solution for theoretical and simulation applications and physical detector design and calibration procedures. The successful implementation underscores the considerable potential of machine learning approaches in detector efficiency analysis and optimization workflows while maintaining the crucial connection between data-driven predictions and physical measurement requirements. The calibration protocol's ability to consistently align theoretical predictions with both simulated and experimental results demonstrates particular value for radiation detection system development and characterization.

A comparison of the experimental data with both the full MC simulation results and the ML predictions, shown in Fig. 4, highlights the key advantage of our approach: while both the simulation and he raw ML model exhibit a similar systematic offset from reality, our framework provides a simple and effective single-point calibration method to correct for it, achieving excellent agreement with experiment. The advantage of our ML approach is speedy, accessibility, and calibratability. It uses MC simulations as a foundation to build a tool that is vastly faster and easier to use for routine applications and can be seamlessly corrected to match physical reality, a process that is far more cumbersome with traditional simulation alone.

Online FEP efficiency calculation platform

The machine learning-based prediction process for detector efficiency curves developed in this study was transformed into an online computational platform[1] to ensure user-friendly accessibility. The web interface, developed using the Shiny library [93] in the R programming language, provides researchers with intuitive access to the model while enabling rapid visualization of the results. This publicly accessible platform, a sample screenshot of which is presented in Fig. 6, represents the practical implementation of our research findings, allowing seamless integration of the predictive model into both academic and industrial workflows.

The platform's design specifically addresses two critical requirements: (1) simplifying complex efficiency curve predictions for non-specialist users through an interactive graphical interface and (2) maintaining the full analytical capability required by experts through customizable input parameters. This dual functionality bridges the gap between advanced machine learning applications and practical detector characterization needs, making the theoretical developments immediately applicable to real-world radiation

[1] https://ml-nuclear-group.shinyapps.io/EfficiencyML/

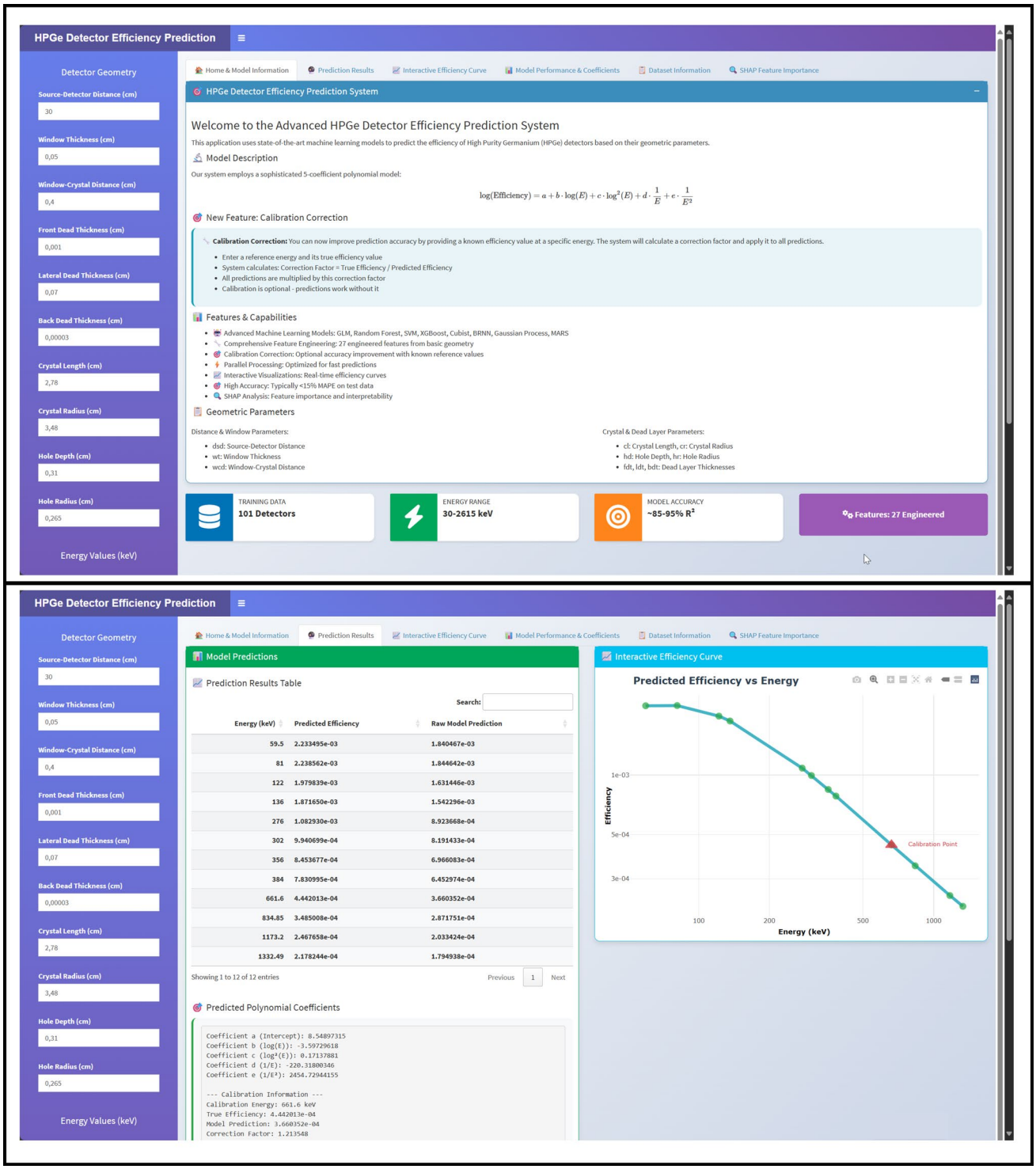


Fig. 6 Screenshots of the online HPGe efficiency prediction platform (https://ml-nuclear-group.shinyapps.io/EfficiencyML/)

measurement. The public availability of this tool ensures that the benefits of the developed methodology can be widely utilized in the nuclear science community.

The developed online applications not only facilitates the effective use of machine learning models in academic and industrial settings but also provides a significant speed advantage over traditional Monte Carlo simulations. While conventional MC simulations can take minutes or even hours, depending on the geometric complexity of the detector, this platform can deliver highly accurate

efficiency predictions within seconds. This feature offers substantial time savings in both the detector design process and laboratory applications.

The computational efficiency stems from the implemented physics-informed machine learning approach, which replaces resource-intensive particle tracking with pretrained models while preserving physical validity through a coefficient-based framework. This makes the tool particularly valuable for applications requiring rapid iteration, such as prototype testing or educational demonstrations, without sacrificing the rigorous physical basis of predictions.

The main interface of the platform (Fig. 6a) accepts the geometric detector parameters (dsd, wt, wcd, fdt, ldt, bdt, cl, cr, hd, and hr) and energy values as the inputs. Users can directly enter these parameters and specify the desired energy points for the efficiency calculations. The system processes these inputs through the previously developed and optimized GLM-based model (detailed in earlier sections) to generate efficiency predictions.

The workflow is intuitive; after entering the required data, users initiate the prediction by clicking the "Predict Efficiency" button. The platform then employs pretrained GLM models to compute and display the results in a tabular and graphical format. The output includes raw model predictions and calibrated predictions (if single-point calibration is applied). The key metrics are the calibration factor, polynomial coefficients, and error reduction rate. The interactive visualization features allow users to dynamically explore the predicted efficiency curves. The calibration point (when used) is highlighted on the graph, enabling an immediate visual assessment of the calibration effects. This functionality particularly supports the rapid analysis of how parameter modifications influence efficiency curves, significantly streamlining detector optimization studies. The platform's responsive design ensures efficient performance, typically generating results within seconds, which is a marked improvement over conventional Monte Carlo approaches. Key advantages of the interface include straightforward parameter input matching detector specifications, instant visualization of calibration impacts, transparent reporting of all prediction components, and interactive tools for sensitivity analysis. This implementation effectively bridges theoretical modeling with practical application requirements while maintaining the physical rigor of the underlying methodology. The public accessibility of this tool further enhances its utility for both research and educational purposes.

This online prediction platform, with its user-friendly interface, high predictive accuracy, and rapid computation capability, serves as a valuable tool for both academic research on HPGe detector parameter sets and industrial applications of HPGe detectors. The platform's implementation is expected to significantly enhance the efficiency of the detector design and calibration processes while enabling the rapid optimization of various detector configurations.

The tool effectively bridges theoretical modeling with practical implementation needs, offering researchers and engineers an accessible and powerful solution for radiation detection system development. Its computational efficiency (i.e., delivering results in seconds rather than hours) makes it particularly valuable for time-sensitive applications and iterative design processes.

Conclusion

This study offers a hybrid machine learning approach for HPGe detector efficiency analysis that transcends the limitations of conventional methods by integrating the interpretability, speed, and precision. Rather than merely developing a predictive model, this study constructed a two-stage framework that performs predictions using physically meaningful polynomial coefficients, as opposed to "black box" methodologies. The high R^2 values (0.975–0.992) achieved using the Generalized Linear Model (GLM) substantiate the theoretical soundness of this approach, while its practical efficacy was unequivocally demonstrated through the single-point calibration protocol. This protocol has served as an effective bridge, diminishing uncertainties in model predictions by an average of 70% and establishing a strong correlation with the experimental reality.

It is important to note that the present model is developed and validated specifically for closed-end coaxial HPGe detectors. The framework itself is generalizable, however, extension to other detector geometries (e.g., BEGe, well-type, or planar) would require the generation of new training dataset encompassing their specific geometric parameters and the re-training of the models.

One of the most important outputs of this study is a publicly accessible online platform that can perform this complex modeling within seconds. This tool democratizes the efficiency calibration process by eliminating the time and expertise barriers required by traditional Monte Carlo simulations. Users obtaining instant reliable predictions for different detector geometries and energy values provides concrete benefits both in accelerating academic research and increasing efficiency in industrial and laboratory applications.

A comparison with established commercial efficiency calibration software, such as LabSOCS or GESPECOR, highlights the complementary strengths and distinct advantages of our proposed framework. While these commercial tools offer robust and validated solutions relying on proprietary efficiency transfer algorithms and requiring precise, often manufacturer-supplied detector characterization files, our physics-informed ML approach provides a transparent

and interpretable alternative. Our model operates solely on fundamental geometric parameters, which reduces vendor dependency and potential associated costs, thereby enhancing accessibility for a wider range of users. Furthermore, while commercial packages typically provide a highly accurate prediction for a nominal detector model, our integrated single-point calibration protocol offers exceptional flexibility. It allows users to swiftly correct the entire efficiency curve to match their specific, real-world detector's performance, a feature that boosts practical utility and streamlines routine laboratory calibrations. The platform delivers reliable results after calibration with a single experimental efficiency measurement, drastically reducing the need for full multi-point calibration.

The proposed framework delivers three key practical outcomes: rapid efficiency prediction (seconds vs. hours for MC), full interpretability meaningful coefficients, and online accessibility through a user-friendly platform. A current limitation is the lack of direct benchmarking against established commercial tools like LabSOCS or GESPECOR, which might be addressed in future work.

In conclusion, this study has the potential to set a new standard for HPGe efficiency curve determination analysis. Enriching the datasets and enhancing the platform's functional capabilities will further strengthen the potential of the proposed method and tool to become a standard reference in applied detector physics. This approach provides a strong and reliable foundation for critical processes such as detector design, optimization, and calibration.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10967-025-10422-6.

Acknowledgements The authors declare that there were no external contributors or funding sources involved in this study. All research presented in this article was completed independently by the authors.

Author contributions CMY: Writing—original draft, Investigation, Methodology, Formal analysis, Data curation, Visualization, Software, Resources; NC: Writing—original draft, Writing—review & editing, Investigation, Software, Resources; AC: Investigation, Resources, Data curation; UC: Writing—review & editing, Investigation, Supervision.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The code and models are written in the R programming environment and are available from the corresponding author upon reasonable requests.

Declarations

Competing interests The authors declare no conflicts of interest. The authors are solely responsible for the content and writing of this paper.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this manuscript, the authors used ChatGPT to improve the English grammar of the text and avoid spelling mistakes. After using this tool/service, the author(s) reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

1. Badawi MS, Gouda MM, Nafee SS et al (2012) New analytical approach to calibrate the coaxial HPGe detectors including correction for source matrix self-attenuation. Appl Radiat Isot 70:2661–2668. https://doi.org/10.1016/j.apradiso.2012.08.014
2. Khan W, Zhang Q, He C, Saleh M (2018) Monte Carlo simulation of the full energy peak efficiency of an HPGe detector. Appl Radiat Isot 131:67–70. https://doi.org/10.1016/j.apradiso.2017.11.018
3. Quang Huy N (2010) The influence of dead layer thickness increase on efficiency decrease for a coaxial HPGe p-type detector. Nucl Instruments Methods Phys Res Sect A Accel Spectrometers Detect Assoc Equip 621:390–394. https://doi.org/10.1016/j.nima.2010.05.007
4. Yang LT, Li HB, Yue Q et al (2019) Search for light weakly-interacting-massive-particle dark matter by annual modulation analysis with a point-contact germanium detector at the China Jinping Underground Laboratory. Phys Rev Lett 123:221301
5. Agostini M, Bakalyarov AM, Balata M et al (2019) Probing Majorana neutrinos with double-β decay. Science 365:1445–1448
6. Aalseth CE, Abgrall N, Aguayo E et al (2018) Search for neutrinoless double-β decay in 76Ge with the Majorana Demonstrator. Phys Rev Lett 120:132502
7. Jakovlevs O, Pchelintsev A, Malgin V et al (2018) Development of miniaturized HPGe spectrometers for unmanned aerial vehicles. J Instrum 13:T06006
8. Moses EI (2010) The National Ignition Facility and the National Ignition Campaign. IEEE Trans Plasma Sci 38:684–689
9. Shaoen J, Ding Y, Liu S et al (2010) Recent inertial confinement fusion experiments and diagnostic techniques on the Shenguang laser facility. Wuli 39:531–542
10. Murphy TJ, Barnes CW, Berggren RR et al (2001) Nuclear diagnostics for the National Ignition facility. Rev Sci Instrum 72:773–779
11. Zhang C, Hu G, Zeng J et al (2022) Re-construction of a HPGe detector precise modeling for efficiency calibration. Appl Radiat Isot 180:110059
12. Sima O, Arnold D, Dovlete C (2001) GESPECOR: a versatile tool in gamma-ray spectrometry. J Radioanal Nucl Chem 248:359–364
13. A.N.G.L.E. 4 (2016) Efficiency calculations for semiconductor and scintillation detectors, user guide
14. Jovanović S, Dlabač A, Mihaljević N, Vukotić P (1997) Angle: a PC-code for semiconductor detector efficiency calculations. J Radioanal Nucl Chem 218:13–20
15. Briesmeister JF (2000) MCNPTM-A general Monte Carlo N-particle transport code Version4C LA-13709-M User's Manual
16. Lépy MC, Bé MM, Piton F (2004) ETNA (Efficiency Transfer for nuclide activity measurement): software for efficiency transfer and coincidence summing corrections in gamma-ray spectrometry
17. Vidmar T (2005) Efftran—a monte carlo efficiency transfer code for gamma-ray spectrometry. Nucl Instruments Methods Phys Res A 550:603–608
18. Nikolic JK, Rajacic M, Todorovic D, Vidmar T (2016) The first experimental test of the MEFFTRAN software on HPGe detector calibration for environmental samples. J Environ Radioact 165:191–196

19. Agostinelli S, Allison J, Amako K et al (2003) GEANT4—a simulation toolkit. Nucl Instrum Methods Phys Res, Sect A 506:250–303
20. Nelson WR, Hirayama H, Rogers DWO (1985) EGS4 code system, user's manual, SLAC-265
21. Kawrakow I, Rogers DWO (2000) The EGSnrc code system, NRC Report PIRS-701
22. Wang Y, Li J, Gu J et al (2015) Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai (China). Appl Soft Comput 35:280–290. https://doi.org/10.1016/J.ASOC.2015.05.047
23. Jeung M, Baek S, Beom J et al (2019) Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. J Hydrol 575:1099–1110. https://doi.org/10.1016/j.jhydrol.2019.05.079
24. Chen G, Wang Y, Li S et al (2018) Spatiotemporal patterns of PM 10 concentrations over China during 2005–2016: a satellite-based estimation using the random forests approach. Environ Pollut 242:605–613. https://doi.org/10.1016/j.envpol.2018.07.012
25. Bayram T, Yeşilkanat CM, Akkoyun S (2023) Applications of different machine learning methods on nuclear charge radius estimations. Phys Scr. https://doi.org/10.1088/1402-4896/ad0434
26. Dey R, Chinnaesakki S, Dhumale MR et al (2023) Estimation of an empirical formula for efficiency of a BEGe type detector using machine learning based algorithm. Radiat Phys Chem 206:110761. https://doi.org/10.1016/j.radphyschem.2023.110761
27. Akkoyun S, Yesilkanat CM, Bayram T (2024) Generation of fusion and fusion-evaporation reaction cross-sections by two-step machine learning methods. Comput Phys Commun 297:109055. https://doi.org/10.1016/j.cpc.2023.109055
28. Yeşilkanat CM (2020) Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos Solitons Fractals 140:110210
29. Mitchell BOJBO (2014) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4:468–481. https://doi.org/10.1002/wcms.1183
30. Ray S (2019) Introduction to Machine Learning and Different types of Machine Learning Algorithms. Proc Int Conf Mach Learn Big Data, Cloud Parallel Comput Trends, Prespectives Prospect Com 2019:35–39
31. Pylypchynets I, Lengyel A, Parlag O et al (2019) Empirical formula for the HPGe-detector efficiency dependence on energy and distance. J Radioanal Nucl Chem 319:1315–1319. https://doi.org/10.1007/s10967-019-06426-8
32. Zheng HL, Tuo XG, Peng SM et al (2018) Determination of Gamma point source efficiency based on a back-propagation neural network. Nucl Sci Tech 29:1–9. https://doi.org/10.1007/s41365-018-0410-4
33. Dinh Chuong H, Huynh Duy Khang N, Thi Ngoc Trang L et al (2024) A comparative study on full-energy peak efficiency calibration methods for HPGe detector: Virtual point detector, curve fitting, and machine learning models. Nucl Instruments Method Phys Res Sect A Accel Spectromet Detect Assoc Equip 1058:168941. https://doi.org/10.1016/j.nima.2023.168941
34. Yu S, He L, Li Y et al (2025) Pulse shape discrimination using a convolutional neural network in an inverted coaxial point contact HPGe detector. Nucl Instruments Method Phys Res Sect A Accel Spectromet Detect Assoc Equip 1072:13–18. https://doi.org/10.1016/j.nima.2025.170219
35. Hossny K, Salama E, Mageed KF et al (2022) Determination of the full energy peak efficiency of HPGe detector for varying soil compositions and Marinelli beaker dimensions using artificial neural networks. J Instrum 17:P11010. https://doi.org/10.1088/1748-0221/17/11/P11010
36. Wang Z, Zhou Z, Xu W et al (2023) Physics informed neural networks for fault severity identification of axial piston pumps. J Manuf Syst 71:421–437. https://doi.org/10.1016/j.jmsy.2023.10.002
37. Nan J, Dai W (2023) Interpretable neural networks with random constructive algorithm. arXiv Prepr arXiv:2307.00185.
38. Caldwell A, Cossavella F, Majorovits B et al (2015) Signal recognition efficiencies of artificial neural-network pulse-shape discrimination in HPGe $0\nu\ \beta\ \beta$ -decay searches. Eur Phys J C. https://doi.org/10.1140/epjc/s10052-015-3573-8
39. Parsa M (2021) A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted Zn–Pb mineral systems of Western Iran. J Geochem Explor 228:106811. https://doi.org/10.1016/j.gexplo.2021.106811
40. Vidmar T (2005) EFFTRAN - A monte carlo efficiency transfer code for gamma-ray spectrometry. Nucl Instrum Methods Phys Res A 550:603–608. https://doi.org/10.1016/j.nima.2005.05.055
41. Ozben CS, Emirhan EM (2009) A hybrid method to determine efficiency curve of HPGe detectors. Appl Radiat Isot 67:1110–1113. https://doi.org/10.1016/j.apradiso.2009.01.085
42. Pan S, Zheng Z, Guo Z, Luo H (2022) An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. J Pet Sci Eng 208:109520. https://doi.org/10.1016/j.petrol.2021.109520
43. Kvalseth TO (1985) Cautionary note about R2. Am Stat 39:279–285. https://doi.org/10.2307/2683704
44. Neuhaus J, McCulloch C (2011) Generalized linear models. WIREs Comput Stat 3:407–413. https://doi.org/10.1002/wics.175
45. Hector A (2015) Generalized linear models for data with non-normal distributions. In: Hector A (ed) The new statistics with R: an introduction for biologists. Oxford University Press, pp 121–140
46. Hilbe JM (2011) Methods of estimation. In: Hilbe JM (ed) Negative binomial regression, 2nd edn. Cambridge University Press, Cambridge, pp 43–60
47. Yin C, Ai M, Chen X, Kong X (2023) Empirical likelihood for generalized linear models with longitudinal data. J Syst Sci Complex 36:2100–2124. https://doi.org/10.1007/s11424-023-2022-2
48. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge
49. Titsias M (2009) Variational learning of inducing variables in sparse Gaussian processes. In: van Dyk D, Welling M (eds) Proceedings of the Twelfth international conference on artificial intelligence and statistics. PMLR, Clearwater Beach, Florida USA, pp 567–574
50. Friedman J (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67. https://doi.org/10.1214/aos/1176347963
51. Kooperberg C (2001) Multivariate adaptive regression splines. In: Encyclopedia of environmetrics
52. Lewis PAW, Stevens JG (1991) Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). J Am Stat Assoc 86(416):864–877. https://doi.org/10.1080/01621459.1991.10475126
53. Sakamoto W (2007) MARS: selecting basis functions and knots with an empirical Bayes method. Comput Stat 22:583–597. https://doi.org/10.1007/s00180-007-0075-7
54. Martinez DL, Shih DT, Chen VCP, Kim SB (2015) A convex version of multivariate adaptive regression splines. Comput Stat Data Anal 81:89–106. https://doi.org/10.1016/j.csda.2014.07.015
55. Quinlan JR (1992) Learning with continuous classes. Aust Jt Conf Artif Intell 92:343–348
56. Quinlan JR (1996) Bagging, boosting, and C4. 5. In: Proceedings of 14th national conference on artificial intelligence/Aaai/Iaai, vol. 1. pp 725–730

57. Fernández-Delgado M, Sirsat MS, Cernadas E et al (2019) An extensive experimental survey of regression methods. Neural Netw 111:11–34
58. Chen J, He Y, Liang Y et al (2024) Estimation of gross calorific value of coal based on the cubist regression model. Sci Rep 14:23176. https://doi.org/10.1038/s41598-024-74469-3
59. Wu Z, Zhu M, Kang Y et al (2021) Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinform 22:bbaa321. https://doi.org/10.1093/bib/bbaa321
60. Turgeman L, May JH, Sciulli R (2017) Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission. Expert Syst Appl 78:376–385
61. Foresee FD, Hagan MT (1997) Gauss-Newton approximation to Bayesian learning. In: Proceedings of International conference on neural networks (ICNN'97). vol.3, pp 1930–1935
62. Burden F, Winkler D (2009) Bayesian Regularization of neural networks BT - Artificial neural networks: methods and applications. In: Livingstone DJ (ed) Humana Press. Totowa, NJ, pp 23–42
63. Mirikitani DT, Nikolaev N (2010) Recursive Bayesian recurrent neural networks for time-series modeling. IEEE Trans Neural Netw 21:262–274. https://doi.org/10.1109/TNN.2009.2036174
64. Troiano M, Nobile E, Mangini F et al (2024) A comparative analysis of the Bayesian regularization and Levenberg–Marquardt training algorithms in neural networks for small datasets: a metrics prediction of neolithic laminar artefacts. Information. https://doi.org/10.3390/info15050270
65. Akkoyun S, Yeşilkanat CM, Bayram T (2024) Machine learning predictions for cross-sections of ^{43,44}Sc radioisotope production by alpha-induced reactions on Ca target. Nucl Instrum Methods Phys Res Sect B Beam Interact Mater Atoms. https://doi.org/10.1016/j.nimb.2024.165293
66. Dong XX, An R, Lu JX, Geng LS (2022) Novel Bayesian neural network based approach for nuclear charge radii. Phys Rev C. https://doi.org/10.1103/PHYSREVC.105.014308
67. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1111/j.1747-0285.2009.00840.x
68. Drucker H, Surges CJC, Kaufman L et al (1997) Support vector regression machines. Adv Neural Inf Process Syst 1:155–161
69. Ibrahem Ahmed Osman A, Najah Ahmed A, Chow MF et al (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J 12:1545–1556. https://doi.org/10.1016/j.asej.2020.11.011
70. Kecman V (Vojislav) (2001) Learning and soft computing : support vector machines, neural networks, and fuzzy logic models. 541
71. Wu J, Liu H, Wei G et al (2019) Flash flood forecasting using support vector regression model in a small mountainous catchment. Water. https://doi.org/10.3390/w11071327
72. Yu PS, Chen ST, Chang IF (2006) Support vector regression for real-time flood stage forecasting. J Hydrol 328:704–716. https://doi.org/10.1016/j.jhydrol.2006.01.021
73. Breiman L (2001) Random forests. Mach Learn 45:5–32
74. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
75. Ho T (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20:832–844. https://doi.org/10.1109/34.709601
76. Yeşilkanat CM, Akkoyun S (2023) Estimation of fission barrier heights for even-even superheavy nuclei using machine learning approaches. J Phys G Nucl Part Phys
77. Efron B (1992) Bootstrap methods: another look at the Jackknife. 569–593. https://doi.org/10.1007/978-1-4612-4380-9_41
78. Yesilkanat CM (2020) Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos Solitons Fractals 140:110210. https://doi.org/10.1016/j.chaos.2020.110210
79. Panov P, Džeroski S (2007) Combining bagging and random subspaces to create better ensembles. Advances in Intelligent data analysis VII. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 118–129
80. Zhang H, Yin S, Chen Y et al (2020) Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. J Clean Prod 273:122858. https://doi.org/10.1016/j.jclepro.2020.122858
81. Friedman J (2001) Greedy Function approximation : a gradient boosting machine author (s): Jerome H, Friedman Source : The Annals of Statistics, Vol . 29 , No . 5 (Oct ., 2001), pp . 1189–1232 Published by : Institute of Mathematical Statistics Stable URL . Ann Stat 29:1189–1232
82. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min 785–794. https://doi.org/10.1145/2939672.2939785
83. Zhu X, Chu J, Wang K et al (2021) Prediction of rockhead using a hybrid N-XGBoost machine learning framework. J Rock Mech Geotech Eng 13:1231–1245. https://doi.org/10.1016/j.jrmge.2021.06.012
84. R Core Team (2024) R: a language and environment for statistical computing. In: R Found. Stat. Comput. http://www.r-project.org/. Accessed date: 18 Aug 2024
85. Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) Kernlab - an S4 package for kernel methods in R. J Stat Softw 11:1–20
86. Wright MN, Ziegler A (2017) Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1–17. https://doi.org/10.18637/jss.v077.i01
87. Chen T, Guestrin C (2016) Extreme gradient boosting [R package xgboost version 1.5.0.2]. In: Proc ACM SIGKDD Int Conf Knowl Discov Data Min 13–17-Augu:785–794. https://doi.org/10.1145/2939672.2939785
88. Rodriguez PP, Gianola D (2022) https://cran.r-project.org/web/packages/brnn/index.html. In: RCran
89. Kuhn M, Quinlan R (2021) Cubist: rule- and instance-based regression modeling, (R package version 0.4.1). https://cran.r-project.org/web/packages/Cubist/index.html. Accessed date: 02 Feb 2023
90. Friedman JH, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22. https://doi.org/10.18637/jss.v033.i01
91. Milborrow S (2023) Earth: multivariate adaptive regression splines (R package version 5.3.2). https://web.archive.org/web/20230510090239/https://cran.r-project.org/web/packages/earth/index.html. Accessed Date: 10 May 2023
92. Varvia P, Räty J, Packalen P (2023) Mgpr: an R package for multivariate Gaussian process regression. SoftwareX 24:101563
93. Shiny (2023) https://cran.r-project.org/web/packages/shiny/index.html
94. Kuhn M, Wing J, Weston S, et al (2020) Caret: classification and regression training. R Packag version 60–86 https//CRANR-project.org/package=caret
95. Kuhn M (2019) The caret package. https://topepo.github.io/caret/index.html. Accessed 24 Mar 2023
96. Wickham H, François R, Henry L, et al (2023) dplyr: A Grammar of Data Manipulation
97. Gupta A, Shareef M, Twisha M et al (2025) Full energy peak efficiency of a clover detector in direct and add-back modes at different source to detector distances: a comparative study. J Radioanal Nucl Chem 334:3459–3469. https://doi.org/10.1007/s10967-025-10081-7
98. Çelik N (2024) A quantitative analysis on the effect of crystal parameters on full energy peak efficiency of a planer and a

coaxial high purity germanium detector for the energy range 46.5 Kev-2.5 Mev. SSRN. https://doi.org/10.2139/ssrn.4839558
99. Çelik N (2025) A quantitative analysis on the effect of crystal parameters on full energy peak efficiency of a coaxial high purity Germanium detector for the energy range 50 keV-2.5 MeV. Radiat Phys Chem 226:112171. https://doi.org/10.1016/j.radphyschem.2024.112171
100. Travar M, Nikolov J, Todorović N et al (2023) Detailed optimization procedure of an HPGe detector using Geant4 toolkit. J Radioanal Nucl Chem 332:817–828. https://doi.org/10.1007/s10967-023-08810-x
101. Sudarshan M, Singh R (1991) Applicability of the analytical functions and semi-empirical formulae to the full energy peak efficiency of large volume HPGe gamma detectors. J Phys D Appl Phys 24:1693–1701. https://doi.org/10.1088/0022-3727/24/10/001
102. Tzika F, Kontogeorgakos D, Vasilopoulou T, Stamatelatos IE (2010) Application of the monte carlo method for the calibration of an in situ gamma spectrometer. Appl Radiat Isot 68:1441–1444. https://doi.org/10.1016/j.apradiso.2009.11.023
103. Lin M, Wang Y, Qin Z (2023) Determination of detection efficiency on HPGe detector for point-like and volumetric samples based on Geant4 simulations. Appl Radiat Isot 200:110989. https://doi.org/10.1016/j.apradiso.2023.110989