Article | Received 13 April 2026; Revised 23 August 2026; Accepted 25 August 2026; Published 16 September 2026
https://doi.org/10.55092/aimat20260010

Machine learning based prediction of luminescence properties of ion-doped Cs_2ZrCl_6 crystalline powders

Lukai Wang*, Shenghong Wang, Longtao Wu, Zhenyu Wang, Zhiyuan Li, Ye Chen and Zungang Wang*

State Key Laboratory of Chemistry for NBC Hazards Protection, Beijing, China

* Correspondence authors; E-mails: wanglukai@sklnbcpc.cn (L.W.); zhigang7991@163.com (Z.W.).

Highlights:

- CMA-ES with 10-fold cross-validation optimizes model hyperparameters.
- Machine learning predicts luminescence properties of Cs_2ZrCl_6 crystalline powders.
- KNR delivers optimal emission wavelength prediction with MSE = 1.4434.
- MLPR delivers optimal decay time prediction with MSE = 0.9234.
- SHAP analysis quantifies feature contributions to luminescence properties.

Abstract: Predicting the luminescence behavior of Cs_2ZrCl_6 crystalline powders with speed and accuracy is a key prerequisite for designing new scintillators aimed at X-ray imaging. Conventional first-principles calculations based on density functional theory can reproduce crystal luminescence, yet they demand substantial computing power and deep materials-domain knowledge, which together translate into high cost, long development cycles, and limited throughput. To sidestep these bottlenecks, our work adopts a machine learning strategy to predict the luminescence properties of Cs_2ZrCl_6. A training dataset was assembled from the literature and expanded through linear interpolation. The hyperparameters of six regression algorithms, including ridge regression (RR), K-nearest neighbor regression (KNR), support vector regression (SVR), random forest regression (RFR), XGBoost regression (XGBR), and multilayer perceptron regression (MLPR), were tuned by coupling the covariance matrix adaptation evolution strategy (CMA-ES) with 10-fold cross-validation. Once trained, the models captured a reliable composition-structure-property link. By comprehensively evaluating three metrics of the coefficient of determination (R^2), mean absolute error (MAE), and mean squared error (MSE), two optimized luminescence property prediction models were identified. The KNR model was selected for emission wavelength prediction, achieving a high R^2 of 0.9999 and a low MSE of 1.4434. Meanwhile, the MLPR model was chosen for decay time prediction, with a high R^2 of 0.9893 and a low MSE of 0.9234. Additionally, the Shapley Additive exPlanations (SHAP) analysis further quantified feature contributions to target properties, offering physically reasonable interpretations. This work demonstrates the promising capacity of machine learning for fast, reliable prediction of crystal luminescence performance, providing practical support for accelerating materials-oriented research and development.

Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

Keywords: machine learning; scintillation crystal; algorithm; luminescent property; prediction model

1. Introduction

Scintillator films, possessing the advantages of high sensitivity, high resolution, and environmental stability, have been regarded as ideal materials for X-ray imaging applications [1,2]. Within such a film, the scintillator crystal is the component that ultimately dictates imaging quality. Lead-free metal halide perovskites have become the choice of scintillator crystals because they combine excellent luminescence efficiency, high atomic numbers, tunable band gaps, and benign chemistry [3,4]. As a representative lead-free metal halide crystal, Cs_2ZrCl_6 possesses a zero-dimensional perovskite-derived structure [5]. Its lattice is built from discrete $[ZrCl_6]^{2-}$ octahedra with Cs+ ions occupying the voids, a configuration that produces a highly localized electronic structure. Such localization strengthens electron-phonon coupling and favors the formation of self-trapped excitons, which in turn gives rise to efficient luminescence [6]. In addition, the large Stokes shift between the absorption and emission spectra of Cs_2ZrCl_6 crystals curtails self-absorption losses [7,8], further sharpening their appeal as X-ray scintillators.

Three principal routes are currently used to tune the luminescent output of Cs_2ZrCl_6 crystals: adjusting the stoichiometric ratio [9], ligand engineering [10], and ion doping [11,12]. Among them, ion doping is particularly effective because it can substantially reshape the band gap and the luminescence kinetics, offering a direct handle on emission wavelength and decay time. Doping ions act in two complementary ways. On the one hand, they can shift the band gap and redistribute the energy levels [13], which is conducive to moving the emission wavelength and raising the luminescence efficiency. For instance, Chen *et al.* [14] prepared Cs_2ZrCl_6 phosphors in which Sb^{3+} substitutes for Zr^{4+} inside the $[ZrCl_6]^{2-}$ octahedra. At roughly 4 K, the Sb^{3+}-doped material showed two emission bands—blue at 495 nm (from the $^1P_1 \rightarrow {}^1S_0$ transition) and red at 622 nm (from the $^3P_{2,1,0} \rightarrow {}^1S_0$ transition). Moreover, the intensity ratio of blue to red emission could be regulated by varying the Sb^{3+} concentration. Ahmed *et al.* [15] replaced In^{3+} in $Cs_2AgInCl_6$ with Yb^{3+} or Er^{3+}. Under 300 nm excitation, the Er^{3+} dopant emitted near 1540 nm via a $^4I_{13/2}$–$^4I_{15/2}$ f-band transition, while Yb^{3+} emitted near 990 nm through a $^2F_{5/2}$–$^2F_{7/2}$ f–f de-excitation. On the other hand, dopant ions can alter the electronic structure and the phonon-coupling mechanism, affording fine control over luminescence lifetime and spectral shape. Li *et al.* [16] doped Cr^{3+} into Cs_2ZrCl_6 crystals to tune the lifetime. As the Cr^{3+} content rose from 2.5% to 40%, the decay time fell from 35.01 μs to 24.81 μs, a shortening that they ascribed to non-radiative energy-transfer channels switched on between neighboring Cr^{3+} ions. Similarly, Gui *et al.* [17] introduced Li^+ into Rb_2AgCl_3 nanocrystals to adjust the decay time. At the optimal concentration, the Li+-doped nanocrystals decayed in just 4.6 ns, far faster than the 11.0 ns of pristine Rb_2AgCl_3, a rate well matched to high-speed dynamic X-ray imaging.

Conventional development of doped crystals still leans heavily on a trial-and-error workflow [18], which is built on extensive material-preparation experiments and therefore suffers from low throughput and high cost. First-principles calculations rooted in density functional theory can deliver highly accurate results at substantially improved speed [19–21], but their computational expense and limited scalability remain drawbacks [22–24]. By contrast, machine learning has emerged as the most promising accelerator for the development of ion-doped crystals [25], which excels at constructing the intricate composition-structure-property mappings that steer the rational, target-oriented design of crystalline materials [22,26]. Mansouri Tehrani *et al.* [27] applied machine learning to screen for incompressible,

superhard transition-metal borides. Kirman *et al.* [28] coupled machine learning with high-throughput robotics to synthesize new perovskite single crystals; their model predicted the optimal growth conditions and led to the first synthesis of $(3\text{-PLA})_2PbCl_4$ crystals. Behara *et al.* [26] used machine learning to sort ABO_3-type perovskite structures into cubic, tetragonal, orthorhombic, and rhombohedral classes with 80.3% accuracy. Collectively, these studies show that machine-learning methods are well suited to the performance analysis and optimization of crystalline materials.

This work aims to establish an accurate predictive model for the chemical composition (dopant type and concentration), physical structural parameters (dopant ionic radius, charge, electronegativity, mass, tolerance factor, *etc.*) of Cs_2ZrCl_6 crystalline powders, and their luminescence properties (emission spectrum and decay time) by integrating multiple machine learning algorithms. The hyperparameters of six machine learning models, including ridge regression (RR), K-nearest neighbor regression (KNR), support vector regression (SVR), random forest regression (RFR), XGBoost regression (XGBR), and multilayer perceptron regression (MLPR), were collaboratively searched using the covariance matrix adaptation evolution strategy (CMA-ES) combined with 10-fold cross-validation. The three metrics of coefficient of determination (R^2), mean absolute error (MAE), and mean squared error (MSE) were comprehensively evaluated for the six models to identify the optimal model for predicting the performance of ion-doped Cs_2ZrCl_6 crystalline powders. Furthermore, the Shapley Additive exPlanations (SHAP) were employed to reasonably interpret the effects of each feature component on the emission wavelength and decay time of Cs_2ZrCl_6 crystalline powders.

2. Machine learning process

2.1. Data acquisition

The machine learning workflow is illustrated in Figure 1. To efficiently obtain luminescence performance data of ion-doped Cs_2ZrCl_6 crystalline powders, research papers related to the keywords “Dopant” and “Cs_2ZrCl_6 Crystal” were searched in the Web of Science (WOS) citation index database, and the corresponding articles were downloaded from literature databases such as ACS, Wiley, and Elsevier. After excluding literature involving Cs_2ZrCl_6 single crystals, the literature related to ion-doped Cs_2ZrCl_6 crystalline powders was divided into two categories. One category focused on singly doped Cs_2ZrCl_6 crystalline powders, and the other category focused on multi-ion co-doped Cs_2ZrCl_6 crystalline powders. To avoid synergistic effects of co-doped ions on the luminescence properties of Cs_2ZrCl_6 crystalline powders and to prevent the dataset descriptors from becoming overly complex and increasing the difficulty of machine learning training, this work focuses on collecting data related to the chemical composition, physical structure, and scintillation luminescence properties of single-ion-doped Cs_2ZrCl_6 crystalline powders from the literature. The dataset descriptors are detailed in Table 1, and the collected raw data are presented in Table S1.

Table 1. Descriptors in the dataset and their corresponding explanations.

Descriptor	Type	Unit	Explanation
D	String	/	Six dopant elements: Bi, Cr, Mn, Mo, Sb, and Te
D_conc	Numeric	mol%	Concentration of dopant ions
R	Numeric	Å	Radius of dopant ions
EN	Numeric	/	Electronegativity of dopant ions
Valence	Numeric	/	Valence state of dopant ions
M	Numeric	/	Relative mass of dopant ions
t	Numeric	/	Tolerance factor of ion-doped Cs_2ZrCl_6 crystalline powders
μ	Numeric	/	Octahedral factor of ion-doped Cs_2ZrCl_6 crystalline powders
EW	Numeric	nm	Emission wavelength dominated by dopant ions
τ	Numeric	μs	Decay time dominated by dopant ions

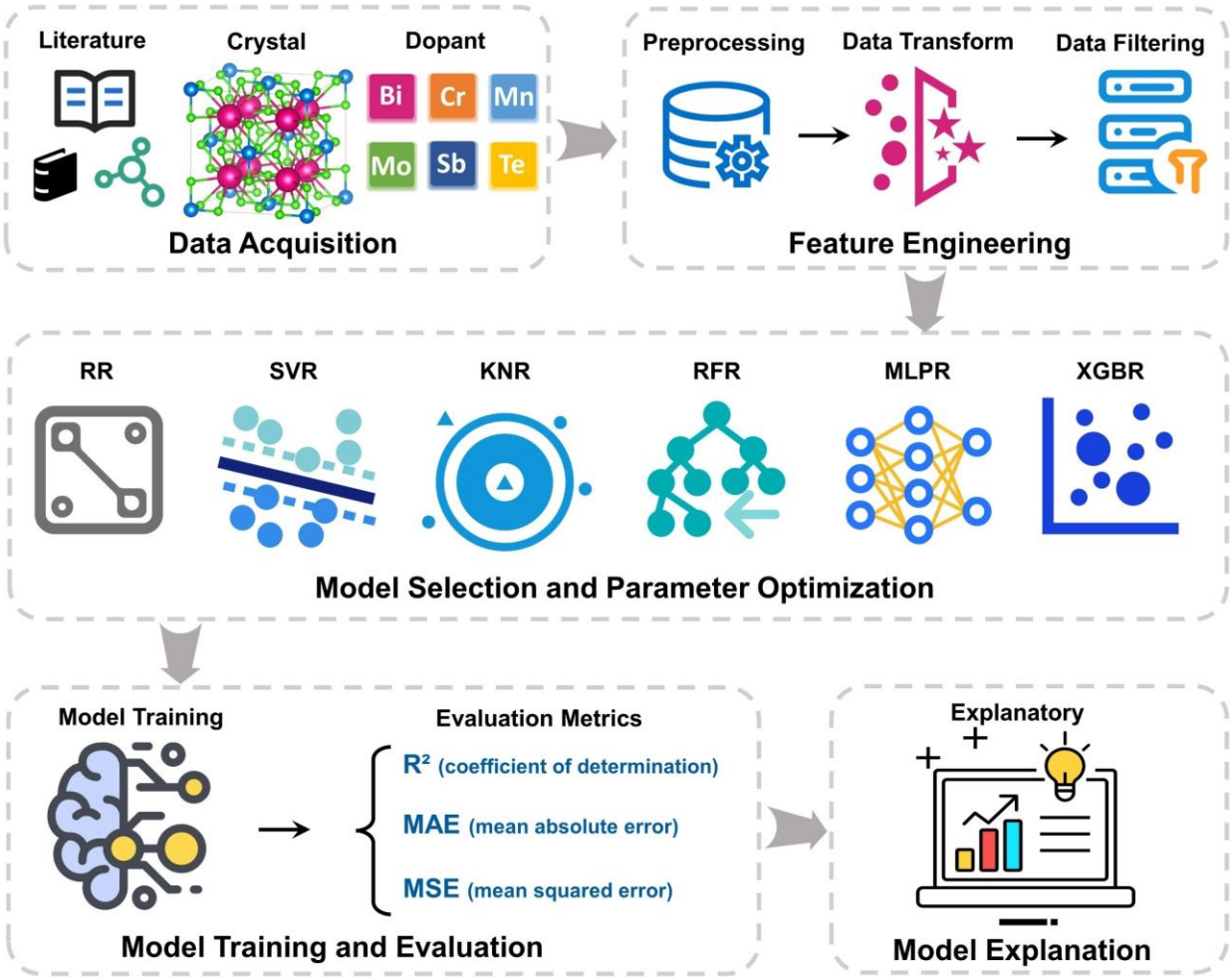

Figure 1. Flowchart of machine learning models for predicting the luminescence properties of ion-doped Cs2ZrCl6 crystalline powders. The graphical elements were provided by AI-generated icons sourced from the iconfont website. The authors reviewed the figure's rationale for explaining machine learning models.

2.2. Feature engineering

The dataset establishment process primarily involves feature engineering. Specifically, it requires completing sample feature selection and transformation during the dataset establishment process.

Dataset construction can be summarized into three steps, including data preprocessing, data transformation, and data screening.

2.2.1. Data preprocessing.

During the data preprocessing stage, the sample descriptors of the dataset must first be determined. Typically, feature variables are derived from known properties of specific elements in the crystal, such as atomic radius, electronegativity, and atomic mass. These feature variables can effectively reflect crystal properties and are sensitive to the target variables, helping to improve model training efficiency and prevent overfitting. Therefore, the number of feature descriptors should be minimized while retaining the essential information of crystals, so as to reduce the dimensionality of the dataset input and the time and resource consumption during machine model training. Based on the collected literature data, six doping elements, including Bi, Cr, Mn, Mo, Sb, and Te, are used to regulate the luminescence properties of Cs_2ZrCl_6 crystalline powders. In this work, the feature descriptors are categorized into two types. The first type consists of basic ionic information, which is obtained by querying relevant databases such as Materials Project, AFLOW, and Wolfram Alpha to retrieve ionic radius, electronegativity, charge state, and other ionic data for the doping elements of Bi, Cr, Mn, Mo, Sb, Te, Cs, Zr, and Cl. The second type involves structural information of the doped Cs_2ZrCl_6 crystalline powders, focusing on two crystal structure parameters involving the tolerance factor and the octahedral factor. The calculation formulas are shown in Equations (1) and (2), respectively.

$$t = \frac{R_A + R_X}{\sqrt{2}\left((1-x)R_B + xR_D + R_X\right)} \quad (1)$$

$$\mu = \frac{(1-x)R_B + xR_D}{R_X} \quad (2)$$

Among these, t represents the Goldschmidt tolerance factor; R_A, R_B, and R_X denote the ionic radii of Cs^+, Zr^{4+}, and Cl^- in Cs_2ZrCl_6 crystalline powders, respectively; R_D represents the ionic radius of the dopant element incorporated into Cs_2ZrCl_6 crystalline powders, and x represents the mole fraction of the dopant ions. In the dataset, the target variables are set as the emission peak wavelength dominated by the dopant element and the corresponding decay time. Considering that most of the literature does not explicitly specify the position of the emission peak wavelength, to enrich the amount of target variable data, GetData software was used to extract the emission peak wavelength data from the emission spectra provided in the literature.

Next, preprocessing operations such as organizing the collected data and handling missing values were carried out. To avoid sample imbalance and excessive weighting of certain features during model training, our work excluded data with the same emission spectrum wavelength or decay time under the same doping element and same doping concentration. At the same time, to prevent the non-uniform distribution of data within the dataset and improve the learning effectiveness of the machine learning model, missing values in the dataset were filled using a linear method. For instance, for the Sb^{3+}-doped Cs_2ZrCl_6 crystal, when the decay time (τ) value was missing at an Sb^{3+} doping concentration of 18 mol%, the corresponding value was filled according to the linear increasing or decreasing trend based on the

decay times at the Sb^{3+} doping concentrations of 15 mol% and 20 mol%. The preprocessed data are detailed in Table S2.

2.2.2. Data transformation

In this work, encoding and transformation are applied to the sample data, particularly non-numerical data, to ensure that the machine learning model can correctly interpret and learn from the dataset [29]. The string-type feature of dopant elements is converted into a set of 0/1 binary features via One-Hot Encoding, where "1" indicates the presence of the dopant element, and "0" indicates its absence. The encoded dataset is further split into a training set and a test set, accounting for 80% and 20% of the total data, respectively.

2.2.3. Data screening

Data screening plays a crucial role in the construction process of datasets, which is a key step to achieve reasonable prediction of the luminescence properties of Cs_2ZrCl_6 crystalline powders. In this study, data screening mainly focuses on sample redundancy and the correlation between features and target variables. Feature correlation analysis methods (e.g., Pearson correlation coefficient [30]) are employed to analyze redundancy among features. For two highly correlated feature variables, the one with stronger predictive power is retained.

2.3. Model selection and hyperparameter optimization

A machine learning model refers to a set of specific linear or nonlinear functions that map input data to output data, establishing relationships between intermediate feature descriptors, target variables, and feature variables. In the machine learning process, an appropriate algorithmic model should be constructed according to the characteristics of the data, and its performance parameters should be adjusted through training. To accurately predict the luminescence properties of Cs_2ZrCl_6 crystalline powders, five supervised learning algorithms from the open-source scikit-learn library and the XGBoost algorithm (Table S3) were selected in this work to construct luminescence performance prediction models. For the different algorithms, CMA-ES combined with 10-fold cross-validation was employed to search for optimal hyperparameters [31,32]. Root Mean Square Error (RMSE), a widely used regression error metric, was adopted to guide the hyperparameter optimization of models. The optimized hyperparameters of six models have been listed in Table S4.

2.4. Model training and evaluation

Six models with optimal hyperparameter configurations were trained. Three metrics, namely the Coefficient of Determination (R^2), MAE, and MSE, were adopted to evaluate the prediction performance of the machine learning models for the luminescence properties of ion-doped Cs_2ZrCl_6 crystalline powders. Based on a comprehensive evaluation, the optimal model was selected, and its predictive advantages were analyzed.

2.5. Model interpretation

The optimal model was interpreted using the SHAP analysis [33]. As an advanced model interpretability tool, SHAP is based on the Shapley value theory from game theory and can provide both global and local explanations for the contribution of each feature to the model prediction results.

SHAP operates on two complementary scales. Globally, it ranks the features that most strongly steer the model's predictions, thereby exposing the dominant patterns and inter-feature couplings embedded in the data. Locally, it dissects individual predictions and shows, feature by feature, how each one pushes the final output up or down.

3. Results and discussion

3.1. Dataset construction and data screening

Relevant literature on ion-doped Cs_2ZrCl_6 crystalline powders was retrieved from the WOS citation database. Following data preprocessing, a raw dataset consisting of 51 samples was obtained (see Table S1). Evidently, the sample size of the current dataset is insufficient to meet the training requirements of machine learning models, which hinders efficient, rapid, and accurate prediction of the luminescence properties of ion-doped cesium zirconium chloride crystals.

Pearson correlation coefficients were then used as the screening metric to weed out redundant variables, thereby simplifying the feature-to-target mapping and improving the efficiency of the subsequent machine-learning training. Figure 2 displays the resulting correlation heatmap, where the color tone and the printed value together encode the strength and direction of the linear association between any two variables. The closer the absolute Pearson coefficient sits to 1, the tighter the linear coupling. Otherwise, the closer it sits to 0, the weaker that coupling.

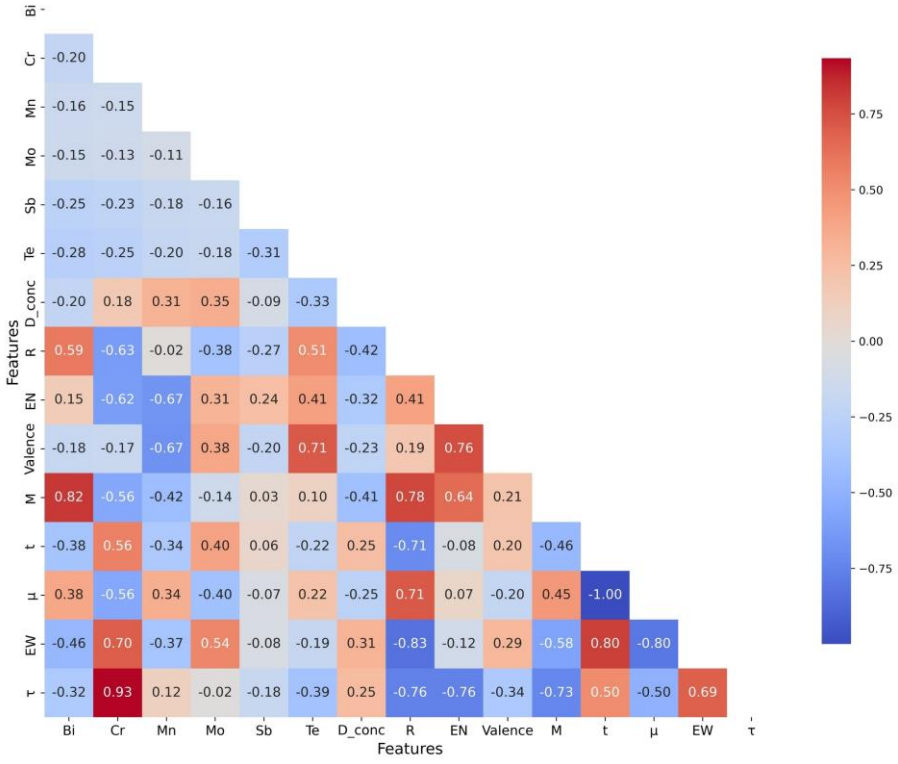

Figure 2. Heatmap of Pearson correlation coefficients for all variables in the dataset.

Examination of the raw dataset (Table S1) reveals that, with varying dopant ion concentrations, the emission wavelength dominated by the same dopant ion remains nearly unchanged, while the decay time shows a regular trend of either increasing or decreasing. Accordingly, linear interpolation was employed in this work to supplement the luminescence performance data of Cs_2ZrCl_6 crystalline powders within specific dopant ion concentration intervals. After data augmentation, the original 51 samples were expanded to 204 samples (see Table S2), which satisfies the data volume demand for machine learning model training.

In Figure 2, the tolerance factor (t) and octahedral factor (μ) exhibit a perfect negative linear correlation, indicating that the two variables can be substituted by a specific mathematical formula. To improve machine learning efficiency, the variable μ was removed from the feature set. In addition, the Pearson correlation coefficient between the Cr dopant ion and decay time is significantly positive, with an absolute value of 0.93, showing a strong positive correlation. This phenomenon to some extent reflects that the Cr dopant ion exerts a strong regulatory effect on the target variable of decay time, providing an important basis for subsequent experimental optimization and theoretical analysis. In view of this, the feature variable of Cr was retained.

3.2. Model evaluation and analysis

Figures 3–5 present the performance comparisons of six machine learning models, namely RR, KNR, SVR, RFR, XGBR, and MLPR, on the two target variables EW and τ. The optimal prediction model was selected by comprehensively evaluating three evaluation indicators (R^2, MAE, and MSE) derived from the independent test set.

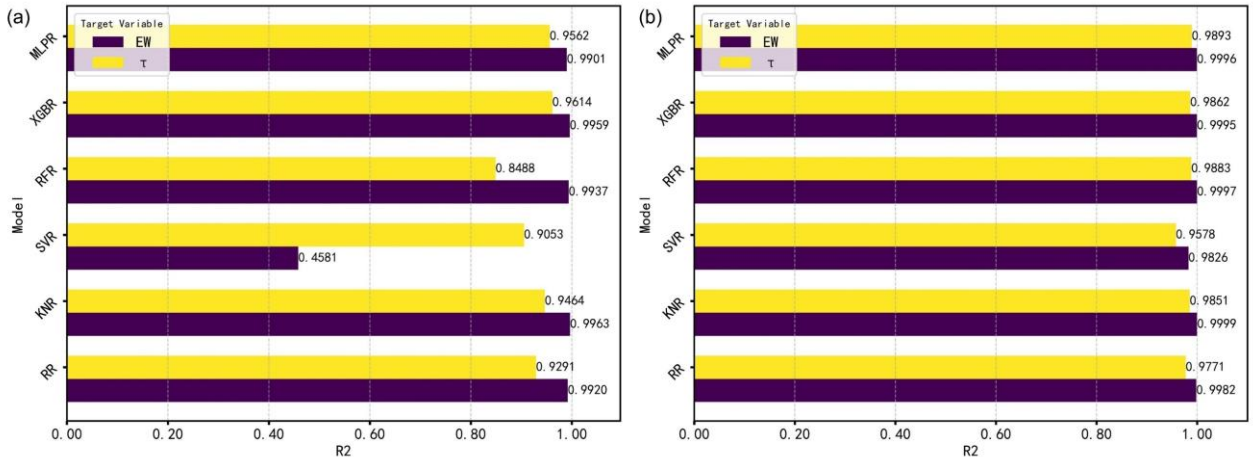

Figure 3. R^2 evaluation metrics of six machine learning models. **(a)** Without and **(b)** With linear interpolation for dataset sample supplementation.

Figure 3 reports the coefficient of determination (R^2) of the six machine-learning models for EW and τ prediction. R^2 spans 0 to 1, with values nearer to 1 reflecting a better goodness of fit. Before linear-interpolation augmentation, the generalization was modest, and SVR was the weakest for EW, returning an R^2 of only 0.4581 (Figure 3a). After augmentation, SVR improved the most, its R^2 jumping to 0.9826 (Figure 3b), and every one of the six models settled near $R^2 = 1.00$ for EW—a sign of excellent predictive power and of a clear EW-to-feature relationship. For τ, augmentation likewise lifted all six

models; RFR gained the most, with R^2 climbing from 0.8488 to 0.9883, while MLPR, XGBR, KNR, and RR all crossed 0.97. MLPR posted the single highest R^2 of 0.9893. SVR's τ score improved from 0.9053 to 0.9578, but it remained the lowest performer, likely because of the intricate distribution of τ or a sub-optimal kernel choice. In all cases, however, augmentation substantially narrowed the performance spread among the algorithms.

To check whether the post-augmentation R^2 gains are simply an overfitting artifact, the training-versus-test performance gap was examined. All six models were evaluated on both EW and τ, and R^2 and MSE were computed for every model-target pair on both data partitions. The R^2 gap (training R^2 minus test R^2) was adopted as the headline overfitting indicator: the larger the positive gap, the weaker the generalization and the more pronounced the overfitting. A three-tier threshold scheme was applied—an R^2 gap above 0.15 signals obvious overfitting with insufficient generalization, a gap in the 0.05–0.15 window marks mild, generally acceptable overfitting, and a gap of 0.05 or below denotes negligible overfitting with strong generalization. The complete numbers are tabulated in Tables S5 and S6 of the Supplementary Data. For both EW and τ, the R^2 gaps of all six models sit comfortably below 0.05, confirming negligible overfitting and solid generalization. Notably, most models exhibit negative R^2 gaps for EW, with test performance surpassing training performance and further eliminating overfitting concerns. These results clearly demonstrate that linear augmentation effectively improves the generalization ability of the machine learning models.

Figure 4 shows the MAE values of different models for EW and τ. A smaller MAE value indicates a more accurate model prediction. For the EW target variable, the SVR model obtains the highest MAE value of 11.9084, followed by RR, indicating that these two models are unsuitable for predicting the EW performance of Cs_2ZrCl_6 crystalline powders. In contrast, the MAE values of KNR, RFR, and MLPR models are all below 2, suggesting higher prediction accuracy for EW. Among them, the KNR model has the smallest MAE value of only 0.7024, demonstrating that the dataset corresponding to EW is more suitable for modeling via KNN regression. For the τ target variable, RFR, KNR, and MLPR models exhibit relatively low MAE values. In particular, the MLPR model has an MAE of only 0.6088, making it more appropriate for predicting the luminescence decay time of Cs_2ZrCl_6 crystalline powders.

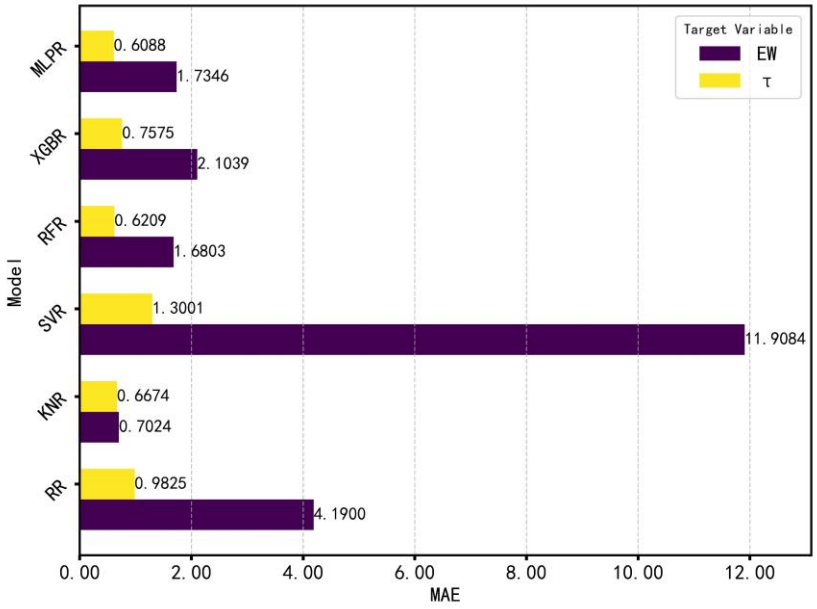

Figure 4. MAE evaluation metrics of six machine learning models.

Figure 5 shows the MSE values of different models for EW and τ. A lower MSE value indicates better prediction performance of the model. For the EW target variable, except for the RR and SVR models, the other four regression models all have small MSE values. Among them, the MSE of KNR is only 1.4434, showing an accurate prediction for EW data. For the τ target variable, the MSE values of the six regression models are all close to 1. The model with the best prediction performance is MLPR, exhibiting the smallest MSE value of 0.9234.

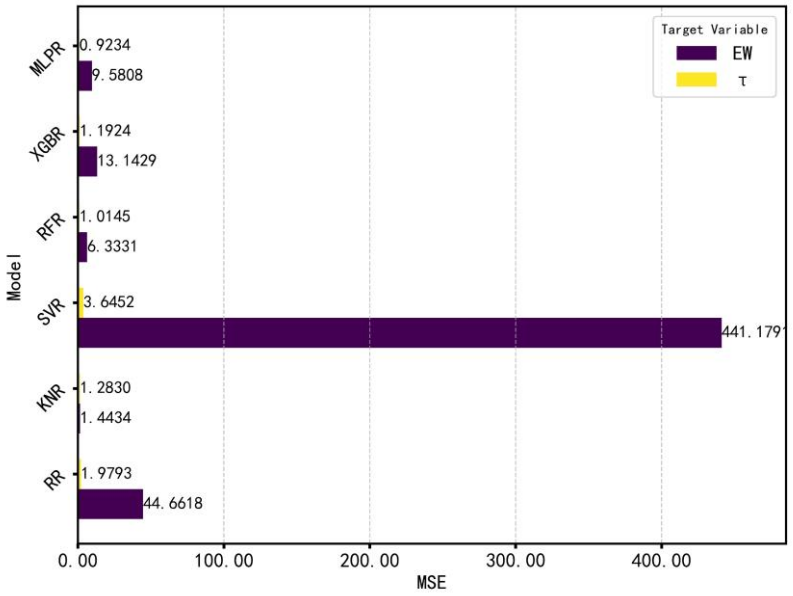

Figure 5. MSE evaluation metrics of six machine learning models.

Figure 6 intuitively presents the prediction results of the six machine learning models for the two target variables of EW and τ. In each scatter plot, the ideal line represents perfect agreement between predicted and actual values. Data points concentrated near the ideal line indicate high consistency and accuracy between model predictions and experimental values. Conversely, data points far from the ideal line imply low prediction accuracy and large errors. Overall, the predicted and actual values of EW and τ are evenly distributed on both sides of the ideal line, confirming that the six regression models deliver relatively good prediction accuracy. Compared with the predicted target values from the test set, those from the training set are closer to the true target values. This phenomenon is mainly related to the fact that the target values in the training set are used for training the machine learning models.

For emission wavelength prediction, all six regression models achieve satisfactory accuracy. Among them, the KNR model performs best, followed by RFR and XGBR, which is consistent with the comprehensive evaluation based on R^2, MAE, and MSE. The SVR model shows the worst prediction performance, with its prediction deviations mainly occurring in ion-doped Cs_2ZrCl_6 crystalline powders characterized by low emission wavelengths. This phenomenon may be caused by the similarity between the emission wavelengths of dopant ions and Cs_2ZrCl_6 hosts, which makes prediction difficult for SVR models. For decay time prediction, the KNR, RFR, XGBR, and MLPR models yield relatively favorable prediction results. Specifically, these four models can accurately predict relatively long decay times but show limited capability in precisely predicting short decay times. Among them, the RFR model is the standout for decay time prediction, accurately estimating the low decay time values of Cs_2ZrCl_6 crystalline powders. By contrast, the RR and SVR models exhibit weak prediction performance for decay time. In particular, the SVR model struggles to accurately predict both high and low decay time values,

showing obvious deviations from the ideal true values. This phenomenon is mainly attributed to the inability of the SVR algorithm to find the optimal separating hyperplane in the feature space.

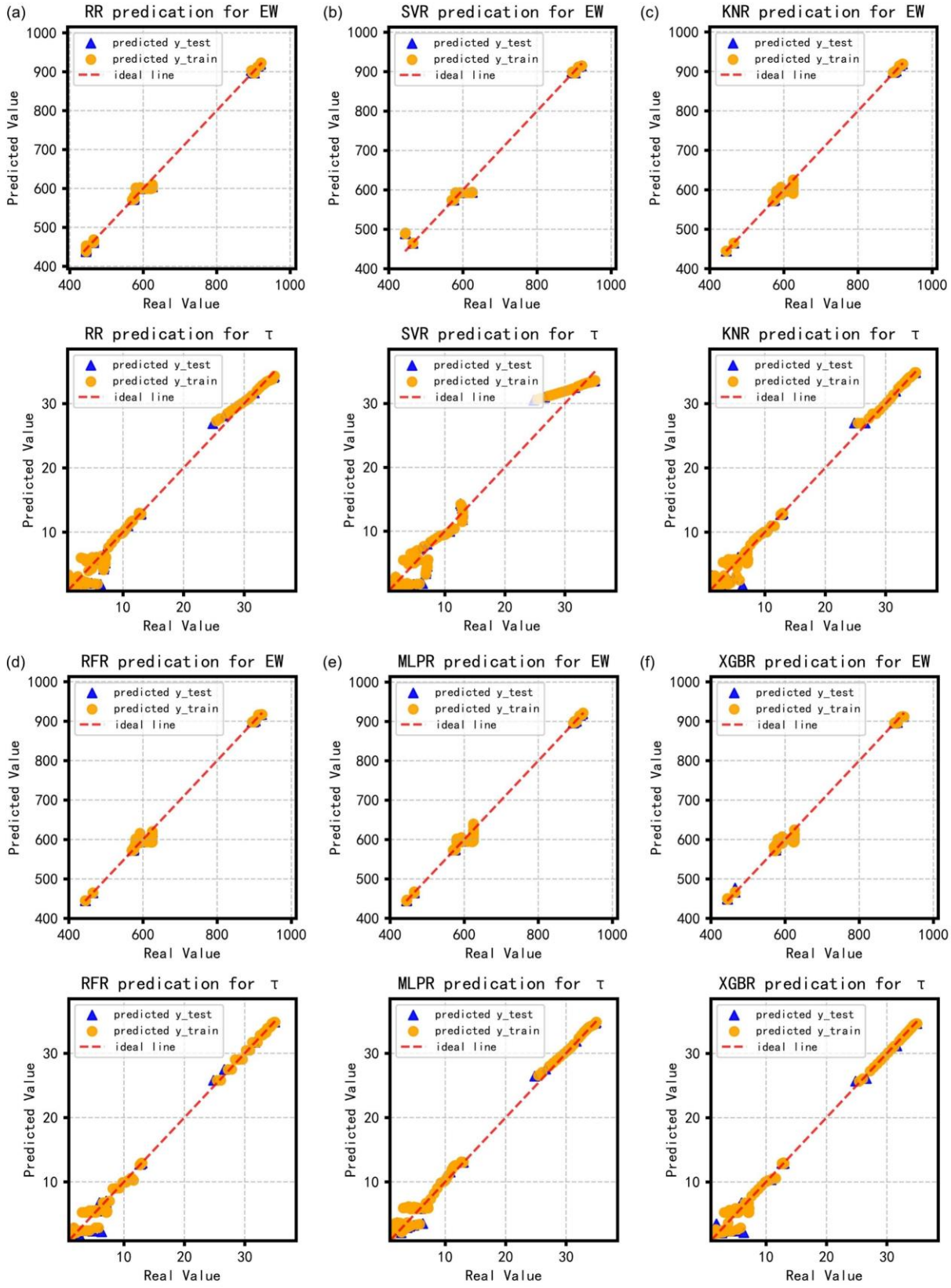

Figure 6. Prediction results for six machine learning models on the two target variables, EW and τ. Predicted *vs.* real values of EW and τ through the **(a)** RR; **(b)** SVR; **(c)** KNR; **(d)** RFR; **(e)** MLPR; and **(f)** XGBR models.

3.3. Interpretation of luminescence performance prediction models

After model evaluation, it is essential to interpret the prediction models to gain in-depth insights into the influence of various feature factors on the luminescence performance of ion-doped Cs_2ZrCl_6 crystalline powders. In this work, SHAP summary plots were used to interpret the KNR model (suitable for EW prediction) and the MLPR model (suitable for τ prediction), respectively. The detailed effects of the 12 input features on the prediction results of the luminescence performance models for ion-doped Cs_2ZrCl_6 crystalline powders are shown in Figures 7 and 8. In the SHAP summary plots, feature importance is listed in descending order on the y-axis, with the topmost feature on the y-axis having the highest importance. The color on the right represents the feature value. Red color indicates a high feature value, while blue color indicates a low feature value. The horizontal axis represents the positive and negative SHAP values, which indicate the likelihood of predicting a specific target. A larger SHAP value corresponds to a stronger positive influence of the feature on the model output, whereas a smaller SHAP value corresponds to a stronger negative influence.

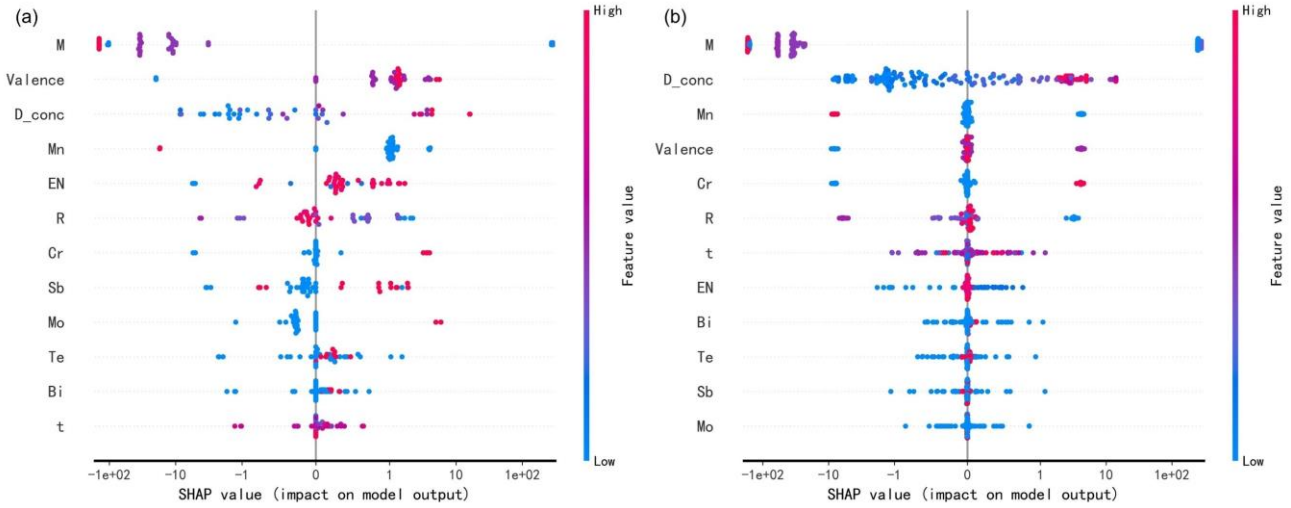

Figure 7. SHAP summary plots of the KNR-predicted dopant-ion-dominated EW performance in Cs_2ZrCl_6 crystalline powders. **(a)** Model trained on the original dataset without linear interpolation; **(b)** Model trained on the linearly augmented dataset.

To identify the features that govern the KNR model's prediction of the dopant-ion-dominated EW in Cs_2ZrCl_6 crystals, the SHAP summary plots for the KNR model trained on the original dataset (without linear interpolation) and on the linearly augmented dataset were compared (Figure 7a,b). Among the 12 input features, dopant ion mass (M) and dopant ion concentration (D_conc) exert the dominant effects on the KNR prediction model for the dopant-ion-dominated EW of Cs_2ZrCl_6 crystals. Specifically, a decrease in M and an increase in D_conc both push the model toward predicting a larger EW. In other words, the KNR model leans toward a larger EW when the dopant is lighter and more concentrated. The effects of M and D_conc on the crystal emission wavelength can be explained from an atomic perspective. When dopant cations replace the B-site elements in Cs_2ZrCl_6 crystals, they affect the structural stability of the lattice. Compared with Zr^{4+} ions, lighter dopant ions such as Mn^{2+} and Cr^{3+} possess smaller ionic radii, which render the Cs_2ZrCl_6 lattice more compact. These structural changes of lattices significantly strengthen the interband interactions, reducing the band gap, and consequently leading to larger EW prediction values. Notably, the directional influence of M and D_conc is virtually

unchanged between the two datasets, confirming that the core physical interpretation is robust and not an artifact of the linear-interpolation augmentation. While the distributions and importance of other features (e.g., Valence, Mn, Cr) show minor variations, these do not affect the dominant role of M and D_conc in determining the dopant-induced EW behavior. Furthermore, a smaller dopant ionic radius (R) in the SHAP values also shows a strong positive influence on the prediction probability of EW values.

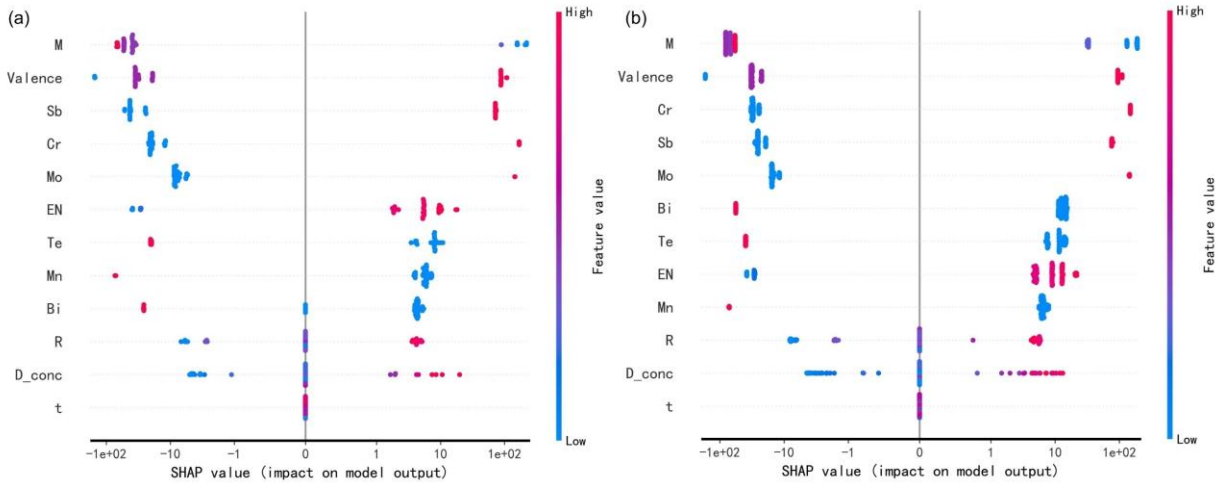

Figure 8. SHAP summary plots of the MLPR-predicted τ performance in Cs_2ZrCl_6 crystalline powders. **(a)** Model trained on the original dataset without linear interpolation; **(b)** Model trained on the linearly augmented dataset.

To map the feature importance that controls the τ prediction of ion-doped Cs_2ZrCl_6 crystalline powders, SHAP analyses were run on the MLPR model trained on both the original and the linearly augmented datasets (Figure 8a, b). Across both scenarios, dopant ion mass (M) and valence state of dopant ions (Valence) consistently emerge as the two most influential features, with M exerting the strongest effect, followed closely by Valence, and this stability across datasets confirms that these two parameters are the core descriptors determining dopant-induced variation in τ, rather than artifacts introduced by linear interpolation. For M, low values correspond to positive SHAP values, while high values correspond to negative SHAP values. This indicates that lighter dopant ions tend to increase the predicted τ, whereas heavier dopants reduce it. This trend arises from atomic-scale effects. Lighter ions introduce less structural strain upon incorporation, which would minimize defect formation and non-radiative recombination pathways to extend carrier lifetime. In contrast, heavier ions may induce greater local lattice distortion, promoting defects that accelerate non-radiative recombination and shorten τ. For Valence, low valence states correspond to negative SHAP values, while high valence states correspond to positive SHAP values. This phenomenon reveals that low-valence dopant ions tend to shorten carrier lifetime τ, whereas higher valence states help elevate the τ value. Such variation originates from charge compensation effects within crystals. High-valence dopants disrupt intrinsic charge balance after occupying lattice positions, which results in the intrinsic defects, including vacancies and interstitials, suppressing non-radiative recombination and effectively extending carrier lifetime. The core feature importance hierarchy and influence trends remain consistent across original and linearly augmented datasets. Linear interpolation exerts no influence on the dominant regulatory effects of M and Valence. Other features, such as R, D_conc, and t, show slight discrepancies in SHAP

distribution. These minor deviations cannot mask the fundamental mechanism dominating τ evolution. All relevant physical interpretations are validated by stable and credible data trends.

4. Conclusion

In summary, a machine learning dataset containing 204 samples was obtained by combining retrieval from the WOS database with a linear interpolation strategy for data imputation. Subsequently, using RMSE as the criterion, CMA-ES coupled with 10-fold cross-validation searched for the best hyperparameters of the six regression models (RR, KNR, SVR, RFR, XGBR, MLPR). The six machine learning models with optimal hyperparameters were trained sequentially to establish the composition-structure-property relationship of crystals. After a comprehensive evaluation of performance metrics, including R^2, MAE, and MSE, the optimal luminescence performance prediction models were identified. Specifically, the KNR model was selected as the optimal choice for predicting the ion-dopant-dominated emission wavelength (EW) of Cs_2ZrCl_6 crystalline powders, achieving excellent performance with a high R^2 of 0.9999 and a low MSE of 1.4434. Meanwhile, the MLPR model exhibited the best performance for decay time (τ) prediction, yielding a high R^2 of 0.9893 and a low MSE of 0.9234. Furthermore, SHAP analysis further quantified how each input feature contributes to the EW and τ predictions, exposing the physical and chemical mechanisms that govern the luminescence of ion-doped Cs_2ZrCl_6 crystalline powders. This work thus confirms that machine learning is a reliable tool for luminescence-property prediction and provides a workable framework for accelerating the development and performance optimization of scintillator crystalline materials.

Data availability statement

The authors confirm that the supplementary data are available within this article, which include raw data collected from the literature (Table S1), data supplemented by interpolation (Table S2), summary of models used in this work (Table S3), summary of optimized hyperparameters used in regression models (Table S4), quantitative overfitting evaluation for the prediction target EW (Table S5), quantitative overfitting evaluation for the prediction target τ (Table S6), and partial code. The relevant code, data, or datasets generated or analyzed in this study are available in ScienceDB at https://www.scidb.cn/anonymous/WkZqRXp1.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used generative AI tools, specifically DeepSeek, Grammarly, and Doubao, for language polishing and text refinement. The partial graphical elements were provided by AI-generated icons sourced from the iconfont website. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work was supported by the Open Fund of State Key Laboratory of Chemistry for NBC Hazards Protection (Grant No. SKLNBC2024-03).

Authors' contribution

Conceptualization, investigation, methodology, writing—original draft, writing—review, editing, and funding acquisition: L.W.; conceptualization, methodology, data curation, writing—review and editing: S.W.; methodology: L.W.; data curation: Z.W.; conceptualization, methodology, and writing—review: Z.L.; data curation: Y.C.; conceptualization, methodology, writing—review, and supervision: Z.W. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Zhou F, Li Z, Lan W, Wang Q, Ding L, *et al.* Halide perovskite, a potential scintillator for X-ray detection. *Small Methods* 2020, 4(10):2000506.

[2] Zhou W, Zhu X, Yu J, Mou D, Li H, *et al.* High-quality $Cs_3Cu_2I_5$@PMMA scintillator films assisted by multiprocessing for X-ray imaging. *ACS Appl. Mater. Interfaces* 2023, 15(32):38741–38749.

[3] Zhang X, Yu S, Meng X, Xiao S. A review on lead-free perovskites for X-ray detection and imaging. *Cryst. Res. Technol.* 2023, 58(5):2200232.

[4] Geng X, Chen Y, Li Y, Ren J, Dun G, *et al.* Lead-free halide perovskites for direct X-ray detectors. *Adv. Sci.* 2023, 10(23):2300256.

[5] Liu R, Zhang W, Li G, Liu W. An ultraviolet excitation anti-counterfeiting material of Sb^{3+} doped Cs_2ZrCl_6 vacancy-ordered double perovskite. *Inorg. Chem. Front.* 2021, 8(17):4035–4043.

[6] Gao W, Abbas N, Xin Y, Wang J, Xiao X, *et al.* Lead-free rare-earth based halide double perovskites: from fundamentals, progress to perspectives. *Laser Photonics Rev.* 2025, 19(14):2500113.

[7] Wang S, Wang L, Wu K, Fu L, Sun L, *et al.* Confined growth strategy enabling Cs_2ZrCl_6 nanocomposite crystals with adjustable particle size and high-efficiency luminescence for additive manufacturing of composite scintillators. *Opt. Mater. Express* 2026, 16(7):1983–1994.

[8] Bulyk LI, Wlodarczyk D, Nagorny S, Nahorna V, Wang P, *et al.* Luminescence of the Cs_2ZrCl_6 under high pressure. *Inorg. Chem.* 2025, 64(2):922–930.

[9] Abfalterer A, Shamsi J, Kubicki DJ, Savory CN, Xiao J, *et al.* Colloidal synthesis and optical properties of perovskite-inspired cesium zirconium halide nanocrystals. *ACS Mater. Lett.* 2020, 2(12):1644–1652.

[10] Zhou L, Fu H, Cao S, Yang W, Zheng J, *et al.* Ligand-mediated kinetic control in coprecipitation: Toward high-performance Sb^{3+}-doped Cs_2ZrCl_6 for single-component white LEDs. *J. Lumin.* 2026, 294:121878.

[11] Zhao D, Liu Y, Han Y, Chen X, Jia M, *et al.* Tunable white light emission in antimony doped vacancy-ordered double perovskite Cs_2ZrCl_6 nanocrystals. *J. Lumin.* 2024, 270:120548.

[12] Shen C, Zhao Y, Yuan L, Ding L, Chen Y, *et al.* Transition metal ion doping perovskite nanocrystals for high luminescence quantum yield. *Chem. Eng. J.* 2020, 382:122868.

[13] Abdelhady AL, Saidaminov MI, Murali B, Adinolfi V, Voznyy O, *et al.* Heterovalent dopant incorporation for bandgap and type engineering of perovskite crystals. *J. Phys. Chem. Lett.* 2016, 7(2):295–301.

[14] Chen C, Xiang J, Chen Y, Jin M, Zheng J, *et al.* White-light emission lead-free perovskite phosphor Cs2ZrCl6: Sb3+. *Ceram. Int.* 2022, 48(2):1851–1856.

[15] Ahmed MS, Sireesha L, Nayak SK, Bakthavatsalam R, Banerjee D, *et al.* Tunable near-infrared emission and three-photon absorption in lanthanide-doped double perovskite nanocrystals. *Nanoscale* 2023, 15(21):9372–9389.

[16] Li X, Yan J, Bezuneh TT, Kou T, Xu J, *et al.* Lead-free halide perovskite Cs_2ZrX_6 doped with Cr^{3+} for multifunctional X-ray and NIR imaging. *J. Phys. Chem. Lett.* 2025, 16(36):9280–9287.

[17] Gui W, Yao L, Zhou X, Wang C. Defect modulation of stable and fast inorganic perovskite Rb2AgCl3 scintillators via Li^+ doping for X-ray imaging. *J. Alloys Compd.* 2025, 1014:178667.

[18] Thakur TS, Dubey R, Desiraju GR. Crystal structure and prediction. *Annu. Rev. Phys. Chem.* 2015, 66:21–42.

[19] Szalewicz K. Determination of structure and properties of molecular crystals from first principles. *Acc. Chem. Res.* 2014, 47(11):3266–3274.

[20] Kang L, Liang F, Jiang X, Lin Z, Chen C. First-principles design and simulations promote the development of nonlinear optical crystals. *Acc. Chem. Res.* 2020, 53(1):209–217.

[21] Podeszwa R, Rice BM, Szalewicz K. Predicting structure of molecular crystals from first principles. *Phys. Rev. Lett.* 2008, 101(11):115503.

[22] Sadeghian M, Palevicius A, Janusas G. A comprehensive review of machine-learning approaches for crystal structure/property prediction. *Crystals* 2025, 15(11):925.

[23] Lu M, Rao S, Yue H, Han J, Wang J. Recent advances in the application of machine learning to crystal behavior and crystallization process control. *Cryst. Growth Des.* 2024, 24(12):5374–5396.

[24] Sun S, Hartono NT, Ren ZD, Oviedo F, Buscemi AM, *et al.* Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* 2019, 3(6):1437–1451.

[25] Ong SP. Accelerating materials science with high-throughput computations and machine learning. *Comput. Mater. Sci.* 2019, 161:143–150.

[26] Behara S, Poonawala T, Thomas T. Crystal structure classification in ABO_3 perovskites via machine learning. *Comput. Mater. Sci.* 2021, 188:110191.

[27] Mansouri Tehrani A, Oliynyk AO, Parry M, Rizvi Z, Couper S, *et al.* Machine learning directed search for ultraincompressible, superhard materials. *J. Am. Chem. Soc.* 2018, 140(31):9844–9853.

[28] Kirman J, Johnston A, Kuntz DA, Askerka M, Gao Y, *et al.* Machine-learning-accelerated perovskite crystallization. *Matter* 2020, 2(4):938–947.

[29] Li S, Liu Y, Chen D, Jiang Y, Nie Z, *et al.* Encoding the atomic structure for machine learning in materials science. *WIREs Comput. Mol. Sci.* 2022, 12(1):e1558.

[30] Jebli I, Belouadha FZ, Kabbaj MI, Tilioua A. Prediction of solar energy guided by pearson correlation using machine learning. *Energy* 2021, 224:120109.

[31] Ikram RM, Goliatt L, Kisi O, Trajkovic S, Shahid S. Covariance matrix adaptation evolution strategy for improving machine learning approaches in streamflow prediction. *Mathematics* 2022, 10(16):2971.

[32] Song Y, Lindsay J, Zhao Y, Nasiri A, Louis SY, *et al.* Machine learning based prediction of noncentrosymmetric crystal materials. *Comput. Mater. Sci.* 2020, 183:109792.

[33] Yin Y, Wang A, Sun Z, Xin C, Jin G. Machine learning regression model for predicting the band gap of multi-elements nonlinear optical crystals. *Comput. Mater. Sci.* 2024, 242:113109.