RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2026, **16**, 11415

Composition-based machine learning for predicting and designing Mn^{4+}-doped phosphors

Ngo T. Que, [a] Vu D. Huan,[b] Le T. Duy,[b] Vu N. Bao,[b] Vu L. Minh, [c] Mai X. Trang, [d] Anh D. Phan †*[ab] and Pham T. Huy [b]

Received 2nd January 2026
Accepted 20th February 2026

DOI: 10.1039/d6ra00029k

rsc.li/rsc-advances

We present a data-driven approach to predict the excitation wavelength, emission wavelength, and crystal field energy levels ($^4T_1$, $^4T_2$) in Mn^{4+}-doped phosphors based solely on elemental composition. For the first time, we construct the largest and most compherensive experimental dataset of Mn^{4+}-activated phosphors to train and accurately predict the properties without relying on complex structural descriptors. Among several evaluated models, the K-Nearest Neighbors and Extra Trees Regressors achieved the highest accuracy for predicting excitation and emission wavelengths, respectively. Importantly, to evaluate generalization, we test these models on Eu^{3+}-doped systems and achieve high predictive accuracy. An inverse design model is further developed to suggest candidate phosphor compositions for target optical outputs. By avoiding complex descriptors while preserving accuracy and interpretability, this work provides a foundation for theory-informed discovery of luminescent materials.

1. Introduction

Phosphor materials, also referred to as luminescent materials, are solids capable of converting various forms of energy into electromagnetic radiation beyond simple thermal emission.[1] Due to their unique optical properties, phosphors have become increasingly significant in both academic research and practical applications. These materials are extensively employed across diverse applications such as display technologies,[2–4] sensors,[5–7] biomedical imaging,[7] food quality analysis,[8] health monitoring,[9,10] and agriculture.[11] Phosphors typically consist of a host material doped with luminescent activator ions. Among these activators, Mn^{4+} ions are well known as efficient red-light emitters because of their excellent thermal and chemical stability, low cost, and environmental friendliness.[1,2,11] Mn^{4+}-doped phosphors exhibit broadband excitation and sharp red emission lines that primarily arise from the $^2E \rightarrow {}^4A_2$ transition. Their optical properties can be interpreted using the Tanabe–Sugano diagram, which describes electronic transitions from the $^4A_2$ ground state to the $^4T_2$ and $^4T_1$ excited states, as well as the $^2E \rightarrow {}^4A_2$ transition. To fully exploit their potential, it is essential to investigate key spectroscopic and electronic properties that govern their behaviors. In particular, determining the excitation wavelength, emission peak, and electronic transitions ($^4T_1$ and $^4T_2$) is crucial for controlling the color, brightness, efficiency, and stability of these materials. By studying them, researchers can better design phosphors to meet specific requirements in lighting, imaging, sensing, and other advanced applications.

Accurately predicting the excitation wavelength, emission peak, and the $^4T_1$ and $^4T_2$ energy levels is not only crucial for optimizing the performance of phosphor materials, but also fundamental to advancing our theoretical understanding of their luminescent behavior.[12,13] These physical quantities provide insights into the electronic structure and energy transfer mechanisms that govern how materials interact with light. In particular, the $^4T_1$ and $^4T_2$ energy levels are associated with specific electronic transitions of dopant ions, which influence both the position and intensity of emission bands. By analyzing these transitions, researchers can infer the local coordination environment, crystal field strength, and site symmetry of activator ions within the host lattice.[14] This information is essential for selecting suitable host materials and dopants to achieve the desired emission and thermal stability.[15] Similarly, the emission peak indicates the energy of photons released as excited electrons return to lower energy states, while the excitation wavelength represents the energy required to trigger this luminescent process. Knowing these two quantities helps select suitable excitation sources, enhance color quality, and evaluate optical efficiency of phosphor materials.[16] These understanding plays a key role in facilitating fabrication and application by identifying promising material systems prior to synthesis.

[a]*Phenikaa Institute for Advanced Study, Phenikaa University, Hanoi 12116, Vietnam. E-mail: anh.phanduc@phenikaa-uni.edu.vn*

[b]*Faculty of Materials Science and Engineering, Phenikaa School of Engineering, Phenikaa University, Hanoi 12116, Vietnam*

[c]*Faculty of Science, Engineering and Built Environment, School of Information Technology, Deakin University, Australia*

[d]*Phenikaa School of Computing, Phenikaa University, Hanoi 12116, Vietnam*

† Present address: Center for Materials Innovation and Technology, Vin University, Hanoi, Vietnam.

While experimental techniques such as photoluminescence, photoluminescence excitation, time-resolved luminescence, and temperature-dependent emission analyses have been widely used,[1,17–19] they require costly equipment, demanding sample preparation, specialized environments, and time-consuming procedures. These problems limit their use in high-throughput or exploratory studies. In contrast, theoretical methods based on using machine learning (ML) or deep learning (DL) to analyze database give fast and accurate estimation of key optical parameters using only compositional or structural data.[20,21] These computational approaches dramatically reduce time and resources needed to screen and optimize phosphor materials.

Machine learning and deep learning are increasingly being applied to the research and design of luminescent materials, particularly phosphors used in LED technologies.[22–34] These models allow us to predict the emission wavelength,[23,24,27,30–32] thermal quenching temperature,[27,29,31] spectral bandwidth, and quantum yield[31] based on a material's composition and crystal structure. Algorithms including artificial neural networks (ANN),[33] Gradient Boosting Regression,[23,26,30,31] and Random Forest[22–25,34] have shown strong performance in accelerating the discovery and optimization of phosphor materials. Among various dopants, europium (Eu)-doped phosphors are the most widely studied using machine learning because a large amount of experimental data is available for them.[28–31,33,34] In contrast, other dopant systems remain underexplored because of the lack of comprehensive and publicly available data. Another major challenge in this area is the use of many input features, which is typically between 50 and 150, for most machine learning models.[23,26–29,31–34] These features often include detailed information at the atomic level such as atomic structure data,[29] ionic radii,[23,33,34] atomic weights,[32,34] and electronegativity values.[23,24,27,30,31,34] Furthermore, collecting full data for each material is often difficult and takes time.

These challenges raise important questions about how to improve the accuracy and usefulness of machine learning models for designing phosphor materials. (1) Can the excitation wavelength, emission peak, $^4T_1$ and $^4T_2$ energy levels of Mn^{4+}-doped phosphors be accurately predicted using only elemental composition without relying on experimental properties or complex descriptors? (2) Which machine learning algorithms provide the best predictive accuracy for excitation and emission properties of Mn^{4+}-doped phosphors? (3) Are models trained solely on Mn^{4+}-doped compositions transferable to other dopant systems with different luminescent behavior? (4) Lastly, can an inverse-design approach be developed to propose candidate phosphor compositions based on desired excitation and emission wavelengths? Answering these questions will help create more efficient and generalizable machine-learning tools to better discover and design new phosphor materials.

In this work, we address these challenges by developing a data-driven approach to predict and design phosphor materials. We first collect experimental data on Mn^{4+}-doped phosphors and use it to train machine learning models that predict the excitation wavelength, the emission peak, and the wavelength of the $^4T_1$ and $^4T_2$ transition based solely on chemical composition. To evaluate the generalizability of our approach, we apply the trained models to predict the optical properties of Eu^{3+}-doped phosphors. Once reliable forward prediction models are established, we construct an inverse design algorithm to suggest phosphor compositions based on target properties.

2. Theoretical background

Our modeling workflow consists of six main steps as illustrated in Fig. 1. First, a dataset of Mn^{4+}-doped phosphors is collected from peer-reviewed papers and books. Second, the data is pre-processed and transformed into numerical features based on elemental composition. In the third step, six different machine learning algorithms including Extra Trees (ET), Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting (XGB), Support Vector Regression (SVR), and Decision Trees (DT) are trained to predict the excitation wavelength, emission wavelength, and the $^4T_1$ and $^4T_2$ energy levels. The fourth step uses the coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE) to evaluate model performance. In the fifth step, model generalization is tested using an independent dataset of Eu^{3+}-doped phosphors to assess transferability across different dopant systems. Finally, an inverse design model is constructed to propose new phosphor compositions that match user-defined excitation and emission targets.

2.1. Data collection

Two datasets were used to develop and validate the predictive models. The first dataset, constructed for the first time, contains information on 1734 Mn^{4+}-doped phosphors reported in 271 published studies. In all cases, the activator charge state is taken from experimental papers where Mn is explicitly identified as Mn^{4+}. It was collected to train machine learning models to predict excitation wavelength, emission wavelength, and the $^4T_1$ and $^4T_2$ energy levels. The host materials in this dataset consist of three to six elements including 373 ternary, 1089 quaternary, 252 quinary, and 20 senary compositions. The experimental data covered broad ranges: excitation wavelengths from 253 to 500 nm, emission wavelengths from 600 to 731 nm, $^4T_1$ energy levels from 224 to 487 nm, and $^4T_2$ energy levels from 274 to 680 nm. In total, the dataset includes 1734 data points for excitation and emission wavelengths and 1701 data points for $^4T_1$ and $^4T_2$ energy levels. The second dataset, used to test model generalization, consists of 1665 data points for excitation and emission wavelengths of Eu-doped phosphors taken from ref. 31. Following the source experimental literature, the Eu activator in this set is categorized as Eu^{3+}. All datasets used in this study are provided in the SI.

2.2. Feature engineering

After collecting the datasets, the chemical compositions were converted into a normalized input format suitable for the predictive models. Each material was expressed as $A_aB_bC_c\cdots Mn_{mn}$, where A, B, C, and Mn are the constituent elements, and a, b, c, and mn are their atomic percentages. These percentages were calculated by dividing the number of atoms of each

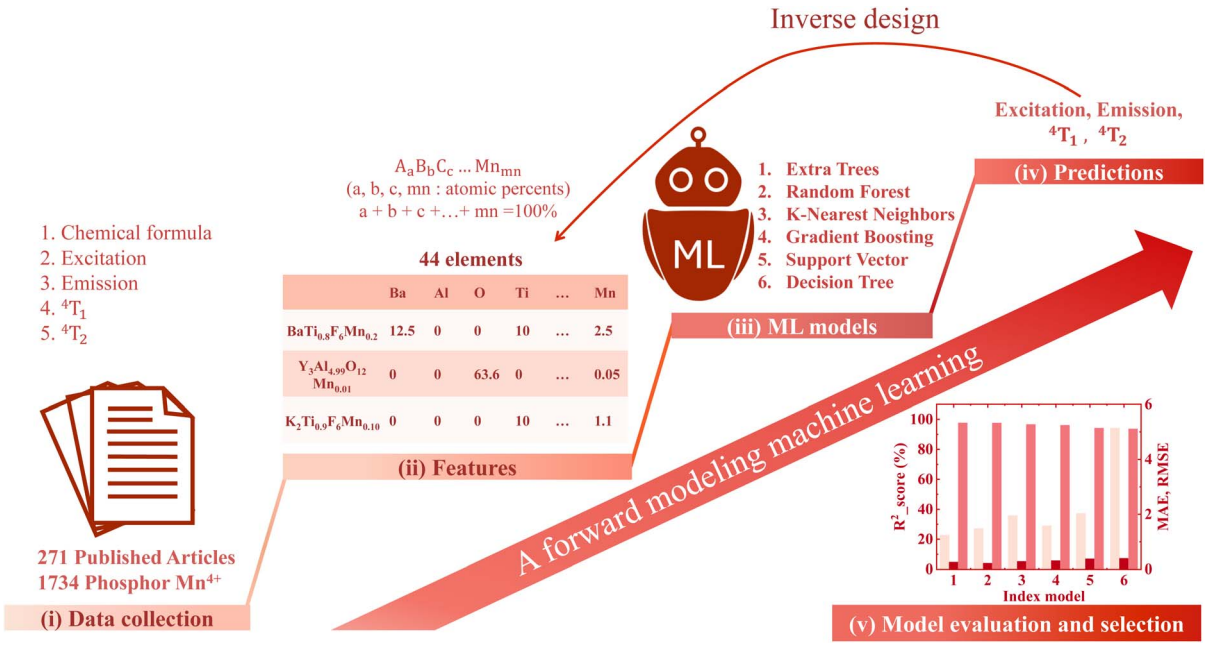

Fig. 1 (Color online) Workflow for predicting the excitation, emission, and energy levels $^4T_1$, $^4T_2$ of Mn^{4+}-activated phosphor using six machine learning algorithms, followed by an inverse design approach to determine the optimal chemical formula of the phosphor material.

element by the total number of atoms in the formula to ensure that the sum equals 100%. The resulting representation is a fixed-length vector comprising 44 features, each corresponding to a possible element, where the feature value is set to zero for elements not present in the composition.

2.3. Machine learning modeling

The dataset was randomly divided into training and testing subsets using an 80 : 20 ratio, which is commonly used in machine learning studies. In our previous work,[21] we examined different splitting ratios ranging from 60 : 40 to 90 : 10 and found that increasing the proportion of training data generally improves the predictive accuracy. However, the improvement becomes marginal when the training set increases from 80 to 90%. To validate our model and avoid overfitting, a 5-fold cross-validation was used during training. Moreover, we performed hyper parameter optimization for each machine learning algorithm using a randomized search approach. In our work, all regression models and the GridSearchCV-based hyper-parameter tuning were implemented using the scikit-learn library.[35,36] This approach scans a broad range of parameter values and selects those that give the best results under cross-validation and improve predictive accuracy.

The model performance was evaluated using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE), defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (2)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (3)$$

where y_i and $\hat{y}_i$ are the observed and predicted values for the ith data point, $\bar{y}$ is the mean of the observed values, and n is the total number of data points. A model with higher R^2 and lower RMSE and MAE values is considered to have better predictive accuracy.

3. Results and discussion

Table 1 provides a summary of previous studies that applied machine learning and deep learning models to predict various properties of phosphor materials. These works include both experimental and DFT-based datasets and use a range of algorithms to model optical and thermal properties such as emission wavelength, thermal quenching temperature, lifetime, and quantum efficiency for different dopant systems. The values in Table 1 serve only as literature benchmarks and are not used as training or test data in our present study. While these studies show the potential of machine learning for phosphor research, they are generally limited by small datasets, narrow dopant types, or the use of complex descriptors that are not always available for new materials. The wide range of predictive accuracy, even when detailed structural or DFT-derived descriptors are employed, indicates that phosphor structure–property relationships remain challenging to capture accurately. There is still significant room for improvement in both descriptors and models. Notably, none of the prior studies analyzed a large and comprehensive dataset for Mn^{4+}-doped phosphors. In this

Table 1 Summary of datasets, ML and DL models, and their corresponding R^2, RMSE and MAE values reported in previous studies for predicting various properties of phosphor materials

Size of data	Type of data	Type of doping	Predicted	DL/ML models	R^2	RMSE	MAE	Reference
39	Experiment	Mn^{4+}	^2E energy (cm^{-1}) (lowest energy excited state)	Linear regression	0.95	149.99	89.33	22
				Robust regression	0.94	153.73	95.68	
				Lasso regression	0.95	149.86	91.05	
				Ridge regression	0.93	168.07	133.81	
				ElasticNet	0.66	383.54	281.9	
				DT	0.31	541.56	401.17	
				RF	0.72	348.04	249.26	
116	Experiment	Mn^{4+}	Emission peak (nm)	XGB	0.71	14.25	9.88	23
				RF	0.80	16.65	10.77	
				Lasso regression	0.64	17.09	11.37	
				Ridge regression	0.69	18.93	12.82	
				KNN	0.85	13.08	8.13	
				SVR	0.81	13.6	9.39	
33	Experiment	Mn^{4+}	Emission peak (nm)	RF	0.87		0.7	24
65	Experiment	Mn^{4+}	Lifetime (ms)	RF			0.432	25
2832	DFT	Ce^{3+}	Relative permittivity (ϵ_r) (eV)	XGB	0.93		0.65	26
219	Experiment	Ce^{3+}	Centroid shift (eV)	XGB	0.90	0.18		
76	Experiment	Ce^{3+}	Emission peak (nm)	Kernel Ridge	0.79		12.64	27
			Thermal quenching (K)		0.64		37	
2610	DFT	Eu^{2+} and Ce^{3+}	Debye temperature (K)	SVR	0.89	59.9	37.9	28
269	Experiment	Eu^{3+}	Thermal quenching (K)	SVR	0.71		31	29
129	Experiment	Eu^{2+}	Emission peak (nm)	XGB	0.78	42		30
1665	Experiment	Eu^{2+} and Eu^{3+}	Emission peak (nm)	XGB	0.866		11.2	31
877			1st excitation max (nm)		0.775		8.83	
951			Decay time (ns)		0.987		0.09	
1252			CIE X coordinate		0.937		0.02	
1252			CIE Y coordinate		0.814		0.02	
183			Thermal quenching (K)		0.574		44.61	
555			Internal quantum efficiency		0.674		9.8	
56			External quantum efficiency		0.675		8.48	
186	Experiment	Cr^{3+}	Emission peak (nm)	SVR	0.821		8.761	32
				KNN	0.85		9.125	
95	Experiment	Eu^{2+}	Excitation wavelength (nm)	CBP	0.999	1.68		33
				Multiple linear	0.999	1.74		
				ANN	0.9999	1.83		
296	Experiment	Eu^{3+}	Asymmetry ratio (Λ)	RF	0.90	1.03	0.77	34

work, we collect such dataset for the first time and use it to develop composition-based machine learning models that accurately predict the excitation wavelength, emission peak, and $^4T_1$ and $^4T_2$ energy levels, as well as to facilitate inverse design of new phosphor compositions.

3.1. Predicting the properties of Mn-doped phosphors

Fig. 2 shows the predictive performance of six machine learning models for estimating the excitation wavelengths of Mn^{4+}-doped phosphors. Among these models, the K-Nearest Neighbors Regressor achieves the best performance with an R^2 of 0.88, an MAE of 8.70 nm, and an RMSE of 21.73 nm on the test set. The remaining models present lower accuracy with R^2 values between 0.77 and 0.86, RMSE values from 23.24 to 29.67 nm, and MAE values from 8.5 to 11.44 nm. These results indicate that KNN captures the relationship between phosphor composition and excitation wavelength more effectively than other models. However, as seen in Fig. 2, there is a smaller subset of samples having noticeably larger deviations. These outliers show that some key factors are not captured by our composition-only representation. In reality, excitation wavelengths are strongly influenced by local coordination geometry, site symmetry, charge-compensation defects, possible multi-centre emission, and experimental issues such as overlapping excitation bands. Because these effects are not explicitly included in our descriptors, these outliers reveal the limits of a purely composition-based model and point to the need for future extensions.

Compared with previous work on Eu-doped phosphors (ref. 31), our models achieve higher R^2 values, which are typically below 0.8 in earlier studies. In contrast, ref. 33 reported R^2 values close to 1 because the dataset is small, less noisy, and based on simple luminescent materials with features strongly related to the predicted property. The evaluation was mainly performed on the training set with a small test set and without rigorous cross-validation. In our case, the Mn^{4+} dataset is much larger and chemically more diverse, so some residual scatter and a limited number of outliers are unavoidable in a minimal-input model. These outliers indicate that additional factors beyond chemical composition such as local structure, defects, or experimental uncertainties also influence the excitation behavior. We therefore view the present results as a realistic

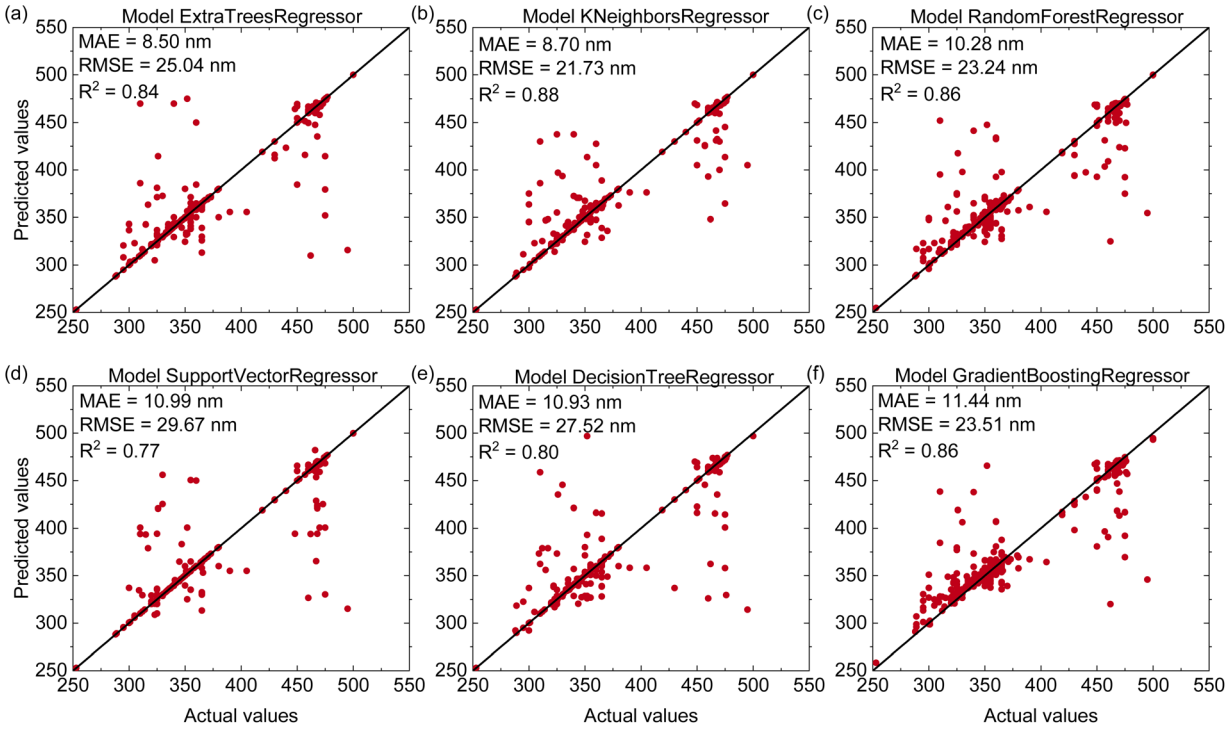

Fig. 2 (Color online) Predictive performance of six regression models for excitation wavelength estimation of Mn^{4+}-doped phosphors on the test dataset including (a) Extra Trees Regressor, (b) K-Nearest Neighbors Regressor, (c) Random Forest Regressor, (d) Support Vector Regressor, (e) Decision Tree Regressor, (f) Gradient Boosting Regressor.

baseline for composition-only predictions and as a starting point for future models that incorporate more detailed structural descriptors.

The emission-wavelength prediction accuracy of six machine learning models for Mn^{4+}-doped phosphors is compared in Fig. 3. All models show high accuracy with R^2 values between 0.94 and 0.98, MAE values ranging from 1.24 to 5.14 nm, and RMSE values from 4.01 to 7.27 nm. Among them, the Extra Trees Regressor exhibits the best performance with the highest R^2 of 0.98, the lowest MAE of 1.24 nm, and RMSE of 4.37 nm. The K-Nearest Neighbors, Random Forest, and Support Vector Regression models also provide good predictions with R^2 values above 0.96. Compared with emission peak predictions in earlier studies[23,24,27,30–32] (Table 1), our models achieve better accuracy due to the larger and more comprehensive dataset and the use of advanced algorithms. The maximum emission wavelengths of Mn^{4+}-doped phosphors typically fall within two ranges: red (620–640 nm) and deep-red/far-red (650–740 nm). The Extra Trees Regressor is the most accurate in the red region, while the Support Vector Regressor performs slightly better for deep-red and far-red emissions. These results suggest that different algorithms capture distinct composition–property relationships, and that ensemble-based methods, particularly Extra Trees, are highly effective for modeling emission behavior.

To gain chemical insight into these predictions, we analyze the feature importance of the Extra Trees model for emission prediction (Fig. S4 in the SI). The analysis shows that fluorine (F), oxygen (O), lanthanum (La), and aluminum (Al) have the highest importance scores, while the remaining elements contribute more weakly. This trend is consistent with physical expectations. The presence of F and O anions significantly affects the local anion environment around Mn^{4+} and therefore have a strong influence on the crystal-field strength, covalency, and nephelauxetic effect.[37] La and Al act as common host cations that control the local coordination geometry and lattice rigidity.[38,39] By contrast, many other cations mainly play secondary structural or charge-balancing roles. As a result, they contribute less independent information to the model and thus receive lower feature-importance scores.

The predictive performance of six machine learning models for estimating the $^4T_1$ energy levels of Mn^{4+}-doped phosphors is shown in Fig. 4. The Decision Tree Regressor achieves the highest accuracy with an R^2 of 0.82, an MAE of 5.44 nm, and an RMSE of 13.04 nm. The remaining models (Extra Trees, Support Vector Regression, K-Nearest Neighbors, Gradient Boosting Regressor and Random Forest) provide slightly lower predictive performance with R^2 values ranging from 0.77 to 0.81. Compared with the emission-peak prediction, the accuracy for $^4T_1$ is clearly reduced. This difference arises mainly from the way $^4T_1$ energies are determined and from their stronger dependence on local structure. The $^4A_2 \rightarrow {}^4T_1$ transition and other electronic transitions such as the charge-transfer band and the $^4A_2 \rightarrow {}^2T_2$ transition can spectrally overlap. This leads to the experimental determination of $^4T_1$ energies less precise and introduces uncertainties into the training dataset. In addition, the $^4T_1$ energy is highly sensitive to local coordination geometry, crystal-field distortions, and covalency, whereas our descriptors do not fully capture these local effects.

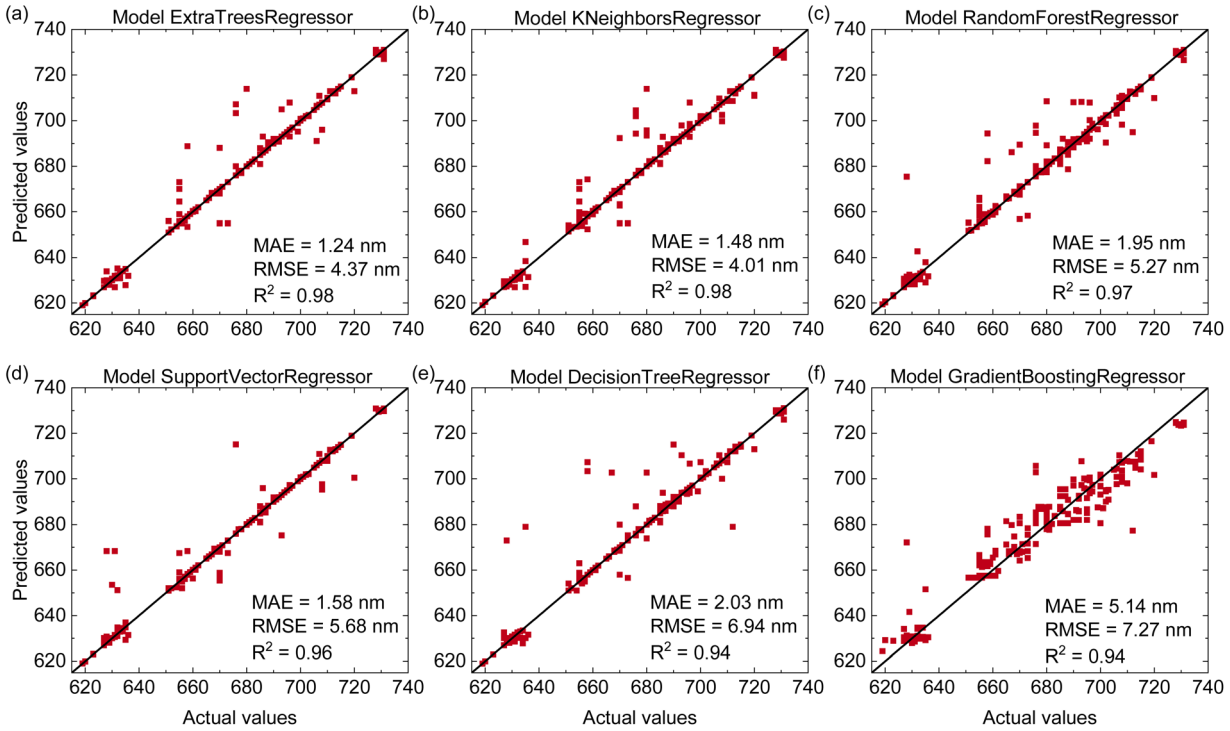

Fig. 3 (Color online) Predictive performance of six regression models for emission-peak estimation of Mn^{4+}-doped phosphors on the test dataset including (a) Extra Trees Regressor, (b) K-Nearest Neighbors Regressor, (c) Random Forest Regressor, (d) Support Vector Regressor, (e) Decision Tree Regressor, (f) Gradient Boosting Regressor.

Fig. 5 shows the predictive performance of the six machine learning models for the $^4T_2$ energy level. Unlike the results obtained for the $^4T_1$ energy level, the K-Nearest Neighbors Regressor outperforms other models with an R^2 of 0.86, MAE of 4.12 nm, and RMSE of 13.96 nm. The Decision Tree Regressor, which previously provided the best results for predicting $^4T_1$

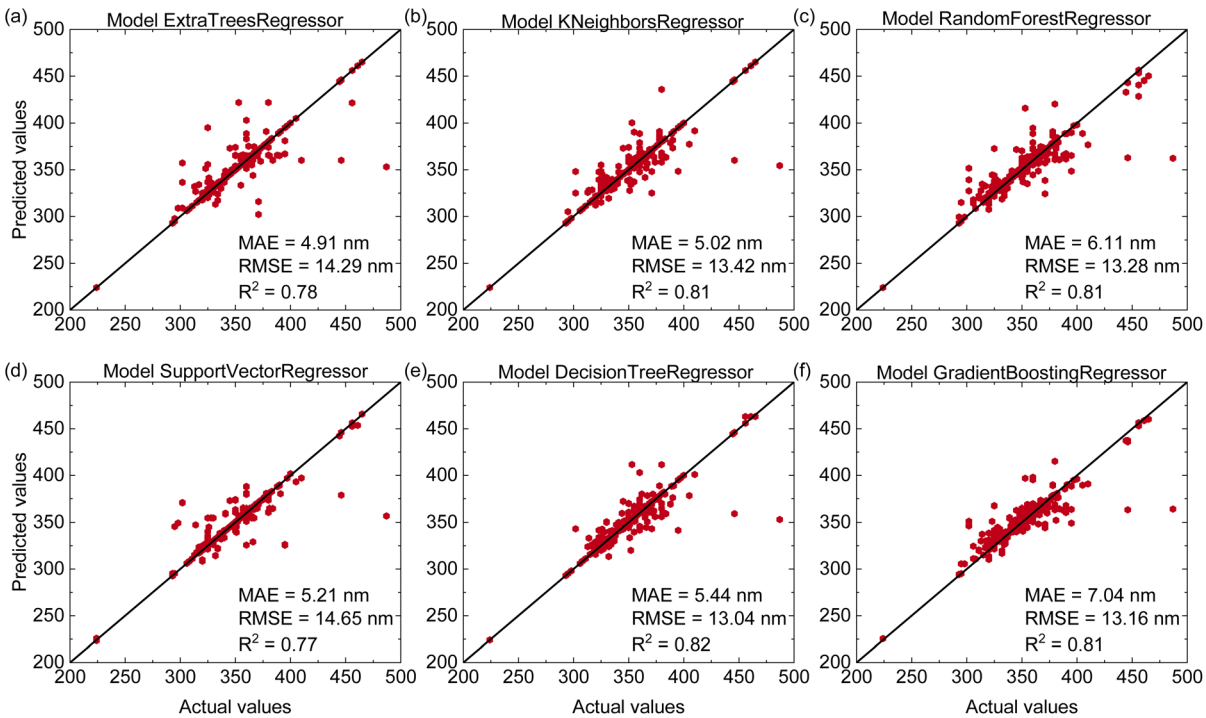

Fig. 4 (Color online) Predictive performance of six regression models for estimating the $^4T_1$ wavelength of Mn^{4+}-doped phosphors on the test dataset including (a) Extra Trees Regressor, (b) K-Nearest Neighbors Regressor, (c) Random Forest Regressor, (d) Support Vector Regressor, (e) Decision Tree Regressor, (f) Gradient Boosting Regressor.

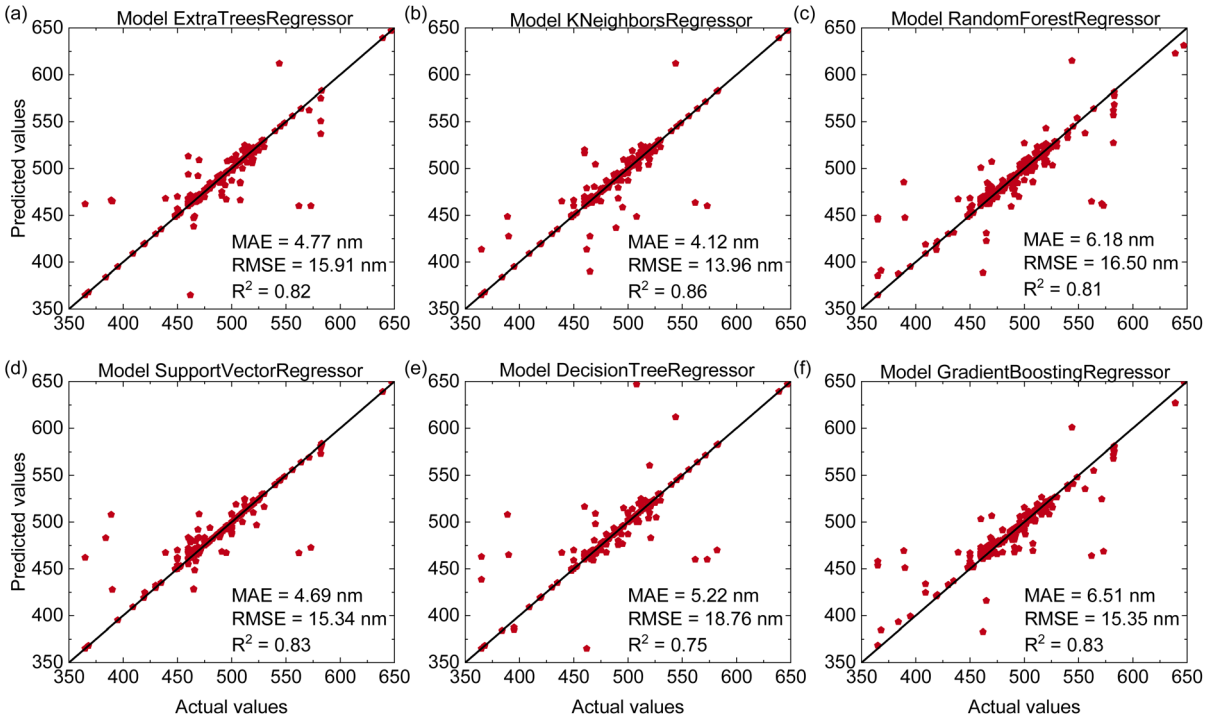

Fig. 5 (Color online) Predictive performance of six regression models for estimating the $^4T_2$ energy levels of Mn^{4+}-doped phosphors on the test dataset including (a) Extra Trees Regressor, (b) K-Nearest Neighbors Regressor, (c) Random Forest Regressor, (d) Support Vector Regressor, (e) Decision Tree Regressor, (f) Gradient Boosting Regressor.

energies, shows significantly lower accuracy for the $^4T_2$ level with an R^2 of 0.75, MAE of 5.22 nm, and RMSE of 18.76 nm. The remaining models present intermediate predictive performance with R^2 values ranging from 0.81 to 0.83. These findings suggest that different electronic transitions exhibit distinct relationships with compositional features and machine-learning models may be particularly effective at modeling the $^4T_2$ energy level.

To further examine the generalization capability of the emission model, we applied the trained Extra Trees regressor to an independent set of Mn^{4+}-doped phosphors that were not used in either training or testing. Specifically, we considered all Mn^{4+}-activated compositions in very recent works[40–47] that (i) exhibit a dominant red emission band and (ii) contain only elements represented in our descriptor space. As shown in Table 2, the absolute differences between predicted and experimental wavelengths range from 1.1 to 26.2 nm, with a mean deviation of approximately 10.7 nm and an RMSE of about 13.2 nm. These errors are larger than the internal test-set, as expected for an external validation set comprising newly reported materials. But these findings indicate that the model is able to provide reasonably accurate first-order estimates of emission peaks for previously unseen Mn^{4+}-doped phosphors. Rather than serving as an exact line-position predictor, the current model is therefore best viewed as a screening tool to identify promising candidate compositions in the desired spectral range.

Table 2 Comparison of emission peak wavelengths (nm) predicted by the Extra Trees regression model with experimental values from recent studies (published in 2025) on Mn^{4+}-doped phosphors

Formula	Actual	Predicted	Ref.
$LaMg_3Sb_{0.999}O_7Mn_{0.001}$	695	696.81	40
$CaYMgNb_{0.997}O_6Mn_{0.003}$	688	691.82	41
$CaAl_2Si_{1.992}O_8Mn_{0.1}$	680	690.9	42
$Mg_{28}Ge_{6.4}Sn_{1.1}O_{32}F_{15.04}Mn_{0.05}$	659	638.37	43
$Ca_{0.8}Na_{0.6}Gd_{0.6}MgWO_6Mn_{0.0005}$	685	698.2	44
$La_3Ga_5Si_{0.9998}O_{14}Mn_{0.0001}$	713	686.8	45
$CsNaWO_2F_4Mn_{0.01}$	631	622.55	46
$Ca_{1.99}Mn_{0.01}La_3Sb_3O_{14}$	709	702.6	47
$Zn_{1.99}Mn_{0.01}La_3Sb_3O_{14}$	690	704.75	47
$Mg_{1.99}Mn_{0.01}La_3Sb_3O_{14}$	705	703.87	47

3.2. Predicting the properties of Eu-doped phosphors

To examine whether a composition-based representation can capture host-dependent trends that extend beyond a single activator ion, we next performed a transferability test on Eu-doped phosphors. Although Mn^{4+} and Eu^{3+} differ in their electronic configurations and detailed emission mechanisms, the positions of their emission bands are physically governed by the host lattice. These host effects are encoded in the elemental composition. Therefore, we apply the same composition-based modeling approach, originally developed for Mn^{4+}-activated phosphors, to an independent Eu^{3+}-doped dataset. This allows us to evaluate whether the approach remains effective across different activator ions.

After training the machine learning models on Mn^{4+}-doped phosphor data, we evaluate their transferability by applying them to a dataset of Eu-doped phosphors. The experimental

dataset was obtained from a recent work of Jang,[31] and the results are presented in Fig. 6. As shown in Fig. 6a, the Extra Trees Regressor predicts the emission peaks with relatively high accuracy, reaching $R^2 = 0.89$, MAE = 7.6 nm, and RMSE = 20.58 nm. It is important to note that the model was trained only on Mn^{4+}-doped phosphors, which emit in the 620–740 nm range, yet it is able to provide reasonably accurate predictions for a broader spectral range of 360–780 nm. In contrast, Fig. 6b presents the excitation wavelength prediction using the Gradient Boosting Regressor, which obtains an R^2 of 0.7, an MAE of 30.03 nm, and an RMSE of 15.88 nm. These results indicate that the cross-dopant excitation predictions are less accurate than the emission predictions.

Compared with the results reported by Jang *et al.*,[31] where emission-peak wavelength prediction achieved $R^2 = 0.866$ and excitation prediction for the first peak reached $R^2 = 0.775$, our model performs competitively or better. Jang's excitation model was trained only on the first excitation peak, while our model was trained on a broader dataset. For a fair comparison, we also retrain our model using only the first excitation peak data and obtain $R^2 = 0.88$. Details of this analysis are provided in the SI. In addition to the validation on Eu-doped phosphors, our approach also shows superior performance on the Mn^{4+}-doped dataset, where the best emission model reaches $R^2 = 0.98$. This indicates a significant improvement over previous models, which reported R^2 values between 0.78 and 0.87.[23,24,27,30–32]

To further validate the generalizability of our model, we compare its predictions with experimental data from several recent studies on Eu^{3+}-doped phosphors published in 2025. Table 3 presents a direct comparison between predicted and experimental emission peaks for a series of compositions not included in the training and testing process. Across these 13 samples, the mean absolute deviation between predicted and experimental values is on the order of 10–20 nm. The Extra Trees model therefore remains reasonable predictive accuracy for Eu^{3+}-doped systems, even though it was trained exclusively on Mn^{4+}-doped data. Using very recent experimental data provides an independent check on model performance and indicates that the proposed framework can be applied to newly reported luminescent materials that were not part of the original training set.

Table 3 Comparison of emission peak wavelengths (nm) predicted by the Extra Trees regression model with experimental values reported in previous studies of Eu^{3+}-doped phosphors

Formula	Actual	Predicted	Ref.
$Sr_3CaNb_{1.994}O_9Eu_{0.06}$	613	613.3	48
$Ca_2MgWO_6Eu_{0.01}Eu_{0.02}$	616	613.64	49
$Y_4Al_2O_9Eu_{0.05}$	611	595.45	50
$La_2LiNbO_6Eu_{0.2}$	613	614.3	51
$LiZnPO_4Eu_{0.03}$	594	572.57	52
$LiSnPO_4Eu_{0.03}$	616	556.17	52
$Sr_4La_6Si_6O_{24}Cl_2Eu_{0.1}$	614	579.35	53
$LaNb_2VO_9Eu_{0.003}$	618	614.95	54
$SrLaZnNbO_6Eu_{0.11}$	618	613.5	55
$Ca_2LaNbO_6Eu_{0.01}$	615	614.15	56
$Ca_3Zr_2SiGa_2O_{12}Eu_{0.1}$	610	572.15	57
$Na_2ZrO_3Eu_{0.002}$	613	602.57	58
$Sr_3La_2W_2O_{12}Eu_{0.09}$	616	608.82	59

3.3. Inverse design

In the final stage of this work, we propose a simple inverse-design approach. Since the Extra Trees Regressor model has been found to have the highest predictive performance for emission wavelength prediction, we choose this model to construct the inverse design framework. The inverse-design calculations use the same datasets as the forward case but the roles of inputs and outputs are reversed. Particularly, the excitation and emission wavelengths are taken as predictors, and the compositional vectors are used as targets. The continuous predictions are then converted into chemical formulas by rounding the atomic fractions to the nearest meaningful values.

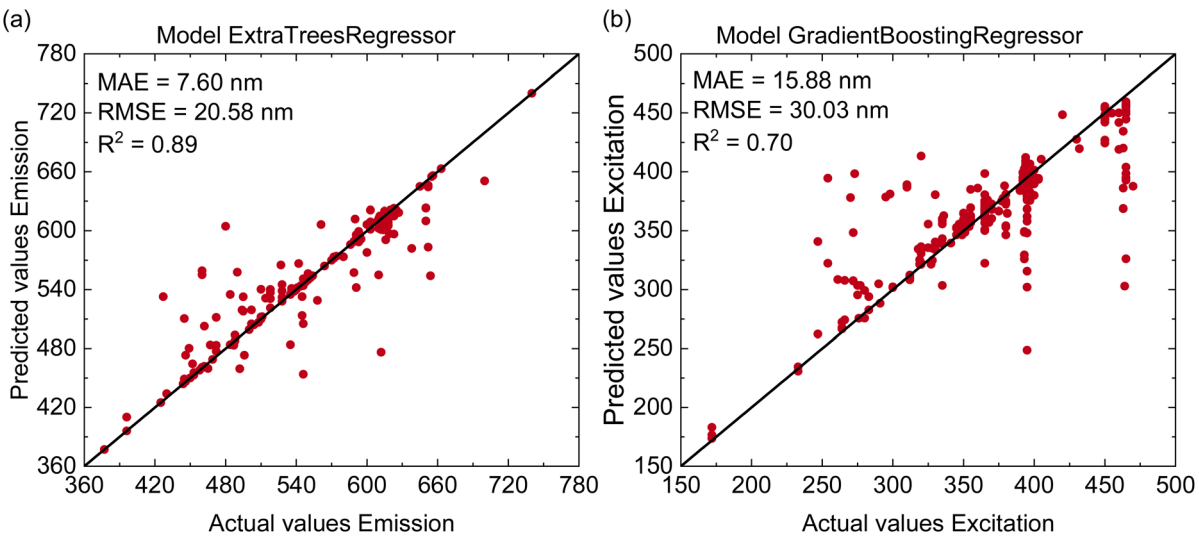

Fig. 6 (Color online) Evaluation of prediction performance for (a) emission peak wavelength using the Extra Trees Regressor and (b) excitation wavelength using the Gradient Boosting Regressor on the test dataset for Eu-doped phosphors.

A predicted composition is regarded as correctly recovered when the reconstructed formula exactly matches the reported experimental one. The continuous predictions are then converted into chemical formulas by rounding the atomic fractions to the nearest meaningful values, and a predicted composition is regarded correctly recovered when the reconstructed formula exactly matches the reported experimental one. To the best of our knowledge, applying a tree-ensemble regressor in this composition-based inverse direction has not previously been proposed for investigating materials.

Our inverse-design approach is then applied to two datasets, one for Mn^{4+}-doped phosphors and one for Eu-doped phosphors. The results show that on the Mn^{4+} dataset, the model successfully predicts the compositions of 265 out of 347 test samples. Similarly, on the Eu dataset, it correctly identifies 144 out of 333 compositions in the test set. Because the Extra Trees regressor is an unconstrained continuous model, its outputs for the atomic fractions are real numbers and are not mathematically forced to satisfy compositional constraints. In principle, the predicted fractions may sum to slightly more or less than 100% or even become negative. In our calculations, we do not observe negative fractions. To obtain chemically meaningful formulas, we therefore discard any predicted composition with a total atomic fraction that deviates from 100%. This screening is applied only to the model outputs in the inverse-design stage. All input compositions in the training and test sets are taken directly from experiment and are already physically valid. Under this constraint, about 96% of the suggested compositions remain valid. This indicates that our inverse-design scheme can propose phosphor compositions from desired optical targets. Rather than serving as a purely generative model, it provides a practical tool to rapidly screen and suggest new phosphor candidates. Thereby, our calculations support experimental synthesis and reduce the time and resources required for materials discovery. The predicted compositions generated by the inverse-design model are listed in SII for Mn^{4+}-doped phosphors and SIII for Eu-doped phosphors in the SI.

Our inverse-design calculations are carried out purely in composition space under simple chemical constraints. All atomic fractions are non-negative, renormalized to sum to 100%, and the Mn or Eu dopant content is limited to the experimental range. The present reverse-engineering scheme operates only at the composition level and is consequently more limited than structure-aware inverse-design approaches that explicitly optimize lattice or microstructural degrees of freedom. However, such structure-resolved methods require reliable crystal structure models and high-cost atomistic calculations. This causes their systematic application to thousands of candidate phosphors to be challenging. Consequently, our inverse-design model is intended as a fast-screening tool that can guide subsequent structure-based simulations or experimental validation.

4. Conclusion

In conclusion, we have developed a composition-based machine learning framework for predicting key optical properties of Mn^{4+}-doped phosphors including the excitation wavelength, emission peak, and $^4T_1$ and $^4T_2$ transition wavelengths. For the first time, we collected the largest experimental dataset of Mn^{4+}-doped phosphors to train and evaluate multiple machine-learning models. Among these models, the K-Nearest Neighbors and Extra Trees Regressors provided the best predictive performance for excitation and emission wavelengths, respectively. The trained models were further validated on Eu^{3+}-doped phosphors to present promising transferability across different dopant systems. An inverse design approach was also developed to generate candidate phosphor compositions based on user-defined optical targets. We note that all Mn-doped compositions in our dataset have nominal Mn contents below 10%, so the predictive performance reported here is valid for the 0–10% Mn-doping range and should not be extrapolated to higher concentrations without additional data. Within the present Mn^{4+} dataset, we do not observe any specific host family or compositional class where the model consistently fails. Prediction errors are distributed across different hosts. Comparable accuracies are also obtained for Eu^{3+}-doped phosphors and in our other phosphor studies (under study and not shown here). These observations indicate that the composition-based approach works reliably across a broad range of chemistries, while still leaving room for future structure-informed models.

This work directly addresses the research questions raised in the Introduction. We showed that accurate predictions of the excitation wavelength, emission peak, and $^4T_1$ and $^4T_2$ transition wavelengths can be determined using only elemental composition without requiring experimental properties or complex descriptors. We further show that models trained solely on Mn^{4+}-doped phosphors can generalize to Eu^{3+}-doped systems, highlighting their transferability across different dopant types. Additionally, our models can be optimized for specific spectral regions and integrated into an inverse design to propose candidate compositions that meet desired optical targets. Compared with previous studies, our approach exhibits higher predictive accuracy while requiring only simple compositional input. Our study provides a minimal-input and data-driven approach for accelerating the discovery and design of high-performance phosphors and expanding the search space for next-generation luminescent materials.

Conflicts of interest

The authors have no conflicts to disclose.

Data availability

The source code used in this study can be found at Github with https://github.com/NgoQue/MLPhosphors.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6ra00029k.

Acknowledgements

This research was funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant No. 103.01-2023.62.

References

1 G. Blasse and B. C. Grabmaier, *Luminescent Materials*, Springer-Verlag, Berlin, Germany, 1994.
2 J. U. Rahman, S. Khan, V. Jain, A. Rajiv, S. Dasi, K. F. Fawy, P. K. Jindal and R. Sivaranjani, *Rev. Inorg. Chem.*, 2025, **45**(1), 55–76.
3 M. Zhao, Q. Zhang and Z. Xia, *Mater. Today*, 2020, **40**, 246–265.
4 H.-W. Chen, J.-H. Lee, B.-Y. Lin, *et al.*, *Light: Sci. Appl.*, 2018, **7**(3), 17168.
5 X. Wang, Q. Liu, Y. Bu, C.-S. Liu, T. Liu and X. Yan, *RSC Adv.*, 2015, **5**(105), 86219–86236.
6 S. E. Crawford, P. R. Ohodnicki and J. P. Baltrus, *J. Mater. Chem. C*, 2020, **8**(24), 7975–8006.
7 J. R. Choi, K. W. Yong, J. Y. Choi, A. Nilghaz, Y. Lin, J. Xu and X. Lu, *Theranostics*, 2018, **8**(4), 1005.
8 C. Wang, X. Wang, Y. Zhou, S. Zhang, C. Li, D. Hu, L. Xu and H. Jiao, *ACS Appl. Electron. Mater.*, 2019, **1**(6), 1046–1053.
9 H. G. Shin, S. Timilsina, K.-S. Sohn and J. S. Kim, *Adv. Sci.*, 2022, **9**(11), 2105889.
10 X. Qian, Z. Cai, M. Su, F. Li, W. Fang, Y. Li, X. Zhou, Q. Li, X. Feng, W. Li, *et al.*, *Adv. Mater.*, 2018, **30**(25), 1800291.
11 M.-H. Fang, Z. Bao, W.-T. Huang and R.-S. Liu, *Chem. Rev.*, 2022, **122**(13), 11474–11513.
12 S. Hariyani, M. Sójka, A. Setlur, *et al.*, *Nat. Rev. Mater.*, 2023, **8**(11), 759–775.
13 X. Zhou, J. Qiao and Z. Xia, *Chem. Mater.*, 2021, **33**(4), 1083–1098.
14 S. Sugano, Y. Tanabe, and H. Kamimura, *Multiplets of Transition-Metal Ions in Crystals*, Academic Press, New York, 1970.
15 Y. Zhou, Q. Ma, M. Lu, Z. Qiu and A. Zhang, *J. Phys. Chem. C*, 2008, **112**(50), 19901–19907.
16 R. S. Yadav and S. B. Rai, *Opt. Laser Technol.*, 2019, **111**, 169–175.
17 I. Pelant and J. Valenta, *Luminescence Spectroscopy of Semiconductors*, Oxford University Press, Oxford, 2012.
18 J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, Springer, 2006.
19 T. H. Gfroerer, *et al.*, *Encycl. Anal. Chem.*, 2000, **67**, 3810.
20 N. T. Que, A. D. Phan, T. Tran, P. T. Huy, M. X. Trang and T. V. Luong, *Mater. Today Commun.*, 2025, **45**, 112287.
21 A. D. Phan, N. T. Que, T. T. Nguyen Duyen, P. Thanh Viet, Q. K. Quach and B. Mei, *J. Appl. Phys.*, 2025, **138**(4), 044703.
22 M. Novita, A. S. Chauhan, R. M. D. Ujianti, D. Marlina, H. Kusumo, M. T. Anwar, M. Piasecki and M. G. Brik, *J. Lumin.*, 2024, **269**, 120476.
23 C. Ding, Z. Li, W. Zhang, J. Ou, X. Wen, C. Xin and M. Su, *New J. Chem.*, 2023, **47**(22), 10875–10883.
24 Y. Wang, W. Tang, C. Zhang, M. S. Molokeev, H. Ming, Y. Zhou, S. Peng, E. Song and Q. Zhang, *Adv. Funct. Mater.*, 2024, **34**(14), 2313490.
25 H. Ming, Y. Zhou, M. S. Molokeev, C. Zhang, L. Huang, Y. Wang, H.-T. Sun, E. Song and Q. Zhang, *ACS Mater. Lett.*, 2024, **6**(5), 1790–1800.
26 Y. Zhuo, S. Hariyani, S. You, P. Dorenbos and J. Brgoch, *J. Appl. Phys.*, 2020, **128**(1), 013104.
27 L. Jiang, X. Jiang, Y. Zhang, C. Wang, P. Liu, G. Lv and Y. Su, *ACS Appl. Mater. Interfaces*, 2022, **14**(13), 15426–15436.
28 Y. Zhuo, A. M. Tehrani, A. O. Oliynyk, A. C. Duke and J. Brgoch, *Nat. Commun.*, 2018, **9**(1), 4377.
29 Y. Zhuo, S. Hariyani, E. Armijo, Z. A. Lawson and J. Brgoch, *ACS Appl. Mater. Interfaces*, 2019, **12**(5), 5244–5250.
30 Y. Koyama, H. Ikeno, M. Harada, S. Funahashi, T. Takeda and N. Hirosaki, *Mater. Adv.*, 2023, **4**(1), 231–239.
31 S. Jang, G. S. Na, Y. Choi and H. Chang, *Sci. Rep.*, 2024, **14**(1), 7639.
32 W. Xu, R. Wang, C. Hu, G. Wen, J. Cui, L. Zheng, Z. Sun, Y. Zhang and Z. Zhang, *npj Comput. Mater.*, 2024, **10**(1), 203.
33 S. K. Sahu, A. Shrivastav, N. K. Swamy, V. Dubey, D. K. Halwar, M. T. Kumar and M. C. Rao, *J. Appl. Spectrosc.*, 2024, **91**(3), 669–677.
34 T. Otsuka, R. Oka, M. Karasuyama and T. Hayakawa, *Phys. Status Solidi RRL*, 2024, **18**(9), 2300237.
35 https://scikit-learn.org/stable/.
36 https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.
37 Y.-I. Kim and P. M. Woodward, *Catalysts*, 2019, **19**(9), 161.
38 Y. Chen, C. Yang, M. Deng, J. He, Y. Xu and Z.-Q. Liu, *Dalton Trans.*, 2019, **48**, 6738–6745.
39 W. Lu, W. Lv, Q. Zhao, M. Jiao, B. Shao and H. You, *Inorg. Chem.*, 2014, **53**(22), 11985–11990.
40 F. Wang and H. Chen, *J. Alloy. Compd.*, 2025, **1020**, 179582.
41 C. Liao, W. Zhao, Y. Xiang, S. Zhu, X. Huang, Z. Chen, J. Xu, C. Jiang, M. Wu and J. Zhong, *Mater. Today Chem.*, 2025, **44**, 102558.
42 F. Chi, J. Zhang, Y. Zheng, X. Niu, J. Liu, X. Zhang, B. Jiang, S. Liu and X. Wei, *Ceram. Int.*, 2025, **51**(2), 2556–2565.
43 M. Li, L. Wang, Q. Shi, H. Guo, J. Qiao, H. Han, C. Cui and P. Huang, *Ceram. Int.*, 2025, **51**(13), 17514–17524.
44 C. Li, R. Kang, X. Ma, J. Xie, Y. Wang and T. Seto, *Small*, 2025, **21**(11), 2500640.
45 N.-N. Zhang, H.-Y. Wang, X.-Y. Yan, X.-P. Wang, B. Liu, Y.-Y. Zhang and Y.-G. Yang, *J. Mol. Struct.*, 2025, **1325**, 141066.
46 W. Fang, Y. Yang, Y. Liu, D. Ma, J. Huang, B. Song and L. Xia, *Inorg. Chem.*, 2025, **64**(8), 4121–4132.
47 H.-Y. Kai, K.-L. Wong and P. A. Tanner, *Next Mater.*, 2025, **8**, 100610.
48 N. Z. Khan, S. A. Khan, N. Muhammad, W. Chen, J. Ahmed, M. A. Padhiar, M. Chen, M. Runowski, S. M. Alshehri and B. Zhang, *Adv. Opt. Mater.*, 2025, **13**(1), 2401938.
49 R. Kiran, S. M. M. Kennedy, A. Princy, M. I. Sayyed, A. H. Almuqrin and S. D. Kamath, *J. Photochem. Photobiol. A Chem.*, 2025, **467**, 116461.
50 R. Arunakumar, M. Gagana, B. R. Radha Krushna, I. S. Pruthviraj, G. Ramakrishna, S. C. Sharma, S. P. N. Choudhury, E. Shanma, B. N. Kumari, K. Manjunatha, *et al.*, *J. Lumin.*, 2025, **281**, 121166.
51 B. Cao, Y. Lu, T. Zhang, H. Wu, Y. Li, C. Deng and W. Huang, *J. Mol. Struct.*, 2025, **1337**, 142193.

52 M. İ. İlhan, L. F. Güleryüz and M. İ. Katı, *Mater. Sci. Eng. B*, 2025, **316**, 118124.
53 Y. Wang, M. Shen, H. Zheng, Y. Lu and P. Du, *Adv. Opt. Mater.*, 2025, **13**(12), 2403214.
54 Y. Du, H. Zhao, T. Guo, Z. Xu, R. Qing, S. Jabeen, J. Che, S. Tong, X. Du and R. Yu, *J. Photochem. Photobiol. A Chem.*, 2025, **468**(1), 116476.
55 M. N. Kumar and P. Samuel, *Ceram. Int.*, 2025, **51**(21), 34819–34830.
56 S. Liu, L. Zhong, Y. Xiang, Z. Chen, M. Xie, J. Hong, L. Zhou and M. Wu, *Mater. Today Chem.*, 2025, **44**, 102556.
57 J. Chen, Y. Chen and H. Guo, *J. Alloy. Compd.*, 2025, **1010**, 177761.
58 P. Khajuria, V. D. Sharma, I. Kumar, A. Khajuria, R. Prakash and R. J. Choudhary, *J. Alloys Compd.*, 2025, **1025**, 180268.
59 F. Wang, H. Chen, S. Zhang and H. Jin, *J. Am. Ceram. Soc.*, 2025, **108**(2), e20200.