

An interpretable linear model bridging data-driven analysis and chemical intuition for Eu^{2+} -phosphor emissions

Yukinori Koyama, Ryusei Hayasaka, Yuta Matsushima, Takayuki Nakanishi, Takashi Takeda & Naoto Hirosaki

To cite this article: Yukinori Koyama, Ryusei Hayasaka, Yuta Matsushima, Takayuki Nakanishi, Takashi Takeda & Naoto Hirosaki (2026) An interpretable linear model bridging data-driven analysis and chemical intuition for Eu^{2+} -phosphor emissions, Science and Technology of Advanced Materials: Methods, 6:1, 2691688, DOI: [10.1080/27660400.2026.2691688](https://doi.org/10.1080/27660400.2026.2691688)

To link to this article: <https://doi.org/10.1080/27660400.2026.2691688>



© 2026 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group



View supplementary material [↗](#)



Published online: 06 Jul 2026.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

An interpretable linear model bridging data-driven analysis and chemical intuition for Eu²⁺-phosphor emissions

Yukinori Koyama^a, Ryusei Hayasaka^b, Yuta Matsushima^b, Takayuki Nakanishi^c, Takashi Takeda^c and Naoto Hiroaki^c

^aCenter for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Japan; ^bDepartment of Applied Chemistry, Chemical Engineering, and Biochemical Engineering, Yamagata University, Yonezawa, Japan; ^cResearch Center for Electronic and Optical Materials, National Institute for Materials Science, Tsukuba, Japan

ABSTRACT

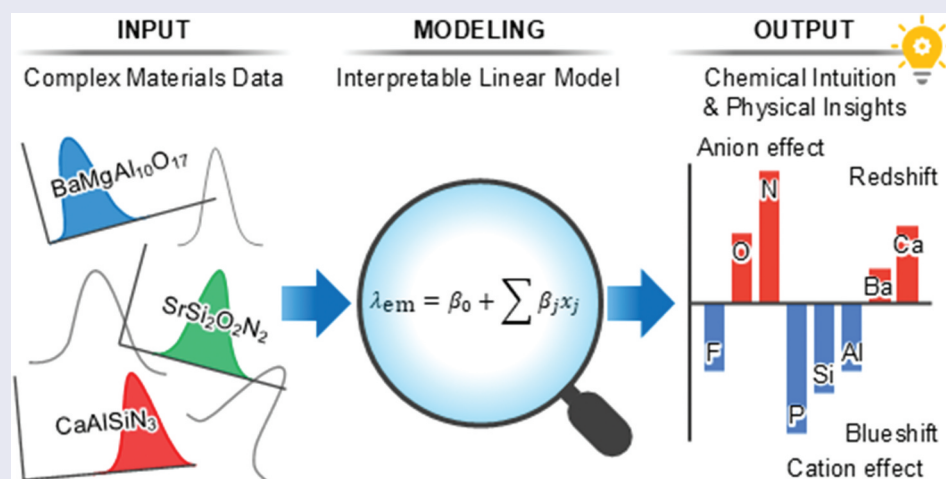
This study demonstrates an interpretable, data-driven framework for analyzing compositional trends in Eu²⁺-activated phosphors. A simple linear model (ridge regression) was constructed using only atomic fractions as features, enabling quantification of the relative influence of each constituent element on the peak emission wavelength through an elemental contribution coefficient (ECC). The ECC is defined as a dataset-conditioned index derived from model coefficients, enabling comparisons of elemental tendencies within the assumptions of the model. The trends observed in the ECC are qualitatively consistent with established empirical rules in phosphor chemistry, thereby supporting the interpretability of the approach. At the same time, systematic discrepancies reveal inherent limitations of composition-based linear models, particularly their inability to separate competing physical effects, such as centroid shift and crystal field splitting. These findings highlight both the potential and the limitations of interpretable models as analytical tools for extracting trends from data and complementing chemical intuition. This study presents a pathway for using machine learning not only as a predictor but also as a tool for exploring data-driven insights in materials science.

ARTICLE HISTORY

Received 24 February 2026
Revised 24 April 2026
Accepted 15 June 2026

KEYWORDS

Materials informatics; phosphor; europium; machine learning; interpretability; linear model; composition-based feature



IMPACT STATEMENT

We propose a simple, interpretable linear model that captures compositional trends in phosphor data in a manner consistent with established chemical understanding, providing a bridge between data-driven analysis and scientific interpretation.

1. Introduction

Phosphor-converted white light-emitting diodes (pc-WLEDs) are central to modern lighting and display technologies due to their high energy efficiency and long lifetime. The performance of these devices,

particularly their color rendering and efficiency, heavily relies on the phosphors that convert excitation light into different colors. Phosphors activated with Eu²⁺ or Ce³⁺ ions have been extensively studied for their high absorption and emission efficiencies, which originate

CONTACT Yukinori Koyama  KOYAMA.Yukinori@nims.go.jp  Data-Driven Inorganic Materials Group, Center for Basic Research on Materials, Sengen 1-2-1, Tsukuba, Ibaraki 305-0047, Japan

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/27660400.2026.2691688>

© 2026 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

from the electric-dipole-allowed 4f–5d transition [1–4]. However, developing new phosphors involves searching for host materials with desired emission properties. This process has traditionally relied on extensive trial-and-error experiments guided by the experience and chemical intuition of researchers, which are time-consuming and costly and often become the rate-limiting step in development.

To overcome this trial-and-error challenge, several semi-empirical physical models have been proposed to quantify chemical intuition. As a pioneering work, Van Uitert reported a simple empirical rule to predict the 5d-level energy from parameters rooted in chemical intuition, namely the valence and coordination number of the luminescence center, the ionic radius of the substituted site, and the electron affinity of the coordinating anion [5]. Subsequently, Dorenbos systematically analyzed over 1,000 inorganic compounds and significantly improved predictive accuracy by modeling the 5d-level energy as a combination of two physical phenomena: centroid shift and crystal field splitting [6]. These models raised chemical intuition into quantitative physical insights, providing indispensable guidance for rational materials design.

In recent years, data-driven machine learning (ML) approaches have emerged as a new trend in phosphor development, driven by advances in materials informatics. Sohn et al. proposed predictive models for the peak emission wavelength using local structural features that inherited the concepts of the physical models [7,8], whereas Nakano et al. used only chemical composition information [9]. Koyama et al. demonstrated the practical utility of ML by discovering new Eu^{2+} -activated green-emitting phosphors based on predictions from an ML model that combined features of both the chemical composition and local structure [10]. A similar success has recently been reported, such as the discovery of a new Eu^{2+} -activated cyan-emitting phosphor by Zhang et al. using ML predictions that combined data from literature and first-principles calculations [11]. Furthermore, Shi et al. constructed a dataset of over 200 phosphors with various luminescence centers, and they showed that their emission wavelengths can be predicted using physicochemical features, such as ionic radii and coordination numbers [12]. The construction and application of these predictive models are supported by the development of comprehensive datasets, such as the one recently built and released by Jang et al., which includes property data for over 2,000 phosphors [13].

These successful studies have established ML as a powerful tool in phosphor discovery. However, many have focused on improving predictive accuracy and discovering new materials, often leading to increasingly complex models. This gives rise to the ‘interpretability dilemma’, a common trade-off between predictive accuracy and model interpretability, which creates

a challenge in various fields of materials informatics [14–17]. Another reason that ML models become ‘black boxes’ in phosphor informatics is the complexity of the features used. For instance, the composition-based features used by Nakano et al. [9] and Koyama et al. [10] were combinations of statistical quantities, such as weighted averages and standard deviations of elemental properties. Although this approach captures the diverse characteristics of host compounds without being limited to specific crystal structures or elemental combinations, it is difficult to quantify the extent to which each element contributes to the prediction. To better understand model predictions, it is important to consider not only ‘what’ a model predicts but also ‘why’ such trends are observed in the data. Therefore, the interpretability dilemma remains a significant challenge in materials informatics.

To address this interpretability dilemma, this study adopts an approach that prioritizes interpretability. Specifically, we limit the features to the atomic fractions of elements, the most fundamental information of a material, and employ a simple linear model (ridge regression) to predict the peak emission wavelength of Eu^{2+} -activated phosphors. The purpose of this intentionally simplified approach is, first, to quantitatively assess the influence of each constituent element on the emission wavelength using the model coefficients. Second, we examine whether the trends extracted from the model are qualitatively consistent with established chemical intuition and empirical knowledge in phosphor chemistry, such as the empirical rules of Van Uitert and Dorenbos. Through this analysis, we explore the potential of using machine learning as an analytical tool for extracting interpretable trends from data, in a manner that complements existing chemical intuition. This study aims to present a framework for analyzing compositional trends transparently, thereby providing a useful basis for hypothesis generation and early-stage screening in materials exploration.

2. Methods

2.1. Data curation and refinement

The dataset of Eu^{2+} -activated phosphors used in this study is based on that constructed in a previous study [10]. Because the reliability of data-driven research depends directly on the quality of the dataset used, we first re-examined and refined the dataset to ensure the transparency and reproducibility of our research. While the original study cited data from a handbook [18] and review articles [1–4], we replaced these citations with the original research papers. In this process, we excluded entries for which the original paper was unavailable from the analysis. Even when the original paper was

accessible, we excluded entries for which the peak wavelength could not be uniquely determined. For instance, we removed entries with multi-peaked emission spectra, particularly when slight changes in relative peak intensities led to a switching of the dominant peak, resulting in discontinuous shifts in the reported peak wavelength. We also note that the emission wavelength of Eu^{2+} -activated phosphors can generally depend on the activator concentration, with variations typically on the order of several to several tens of nanometers depending on the hosts. However, in this study, we did not explicitly consider concentration dependence and instead adopted representative peak wavelengths reported in the literature. This reflects our intention to construct a dataset that captures compositional trends in host materials consistently, while avoiding additional variability arising from differences in reported concentration conditions. Furthermore, we removed data for intermediate compositions of solid-solution compounds and retained only the end-member data.

Following this procedure, we constructed a final dataset consisting of 118 Eu^{2+} -activated phosphors. All host compounds in this dataset contain one alkaline earth metal: Ca, Sr, or Ba. The dataset consists of the following four items and is provided as a comma-separated values (CSV) file in the Supplemental material.

- index: Unique serial number to identify each data entry.
- formula: Chemical formula of the host compound.
- emission: Peak wavelength of the emission spectrum in nanometers (nm).
- doi: Digital object identifier (DOI) of the original research paper from which the data was sourced.

2.2. Feature engineering

In this study, we construct a machine learning model to predict the peak emission wavelength from the chemical composition of the phosphor host. As explanatory variables (features), we used only the atomic fractions of each element in the host compound. This choice inherently imposes a fundamental limitation: The model cannot distinguish between polymorphs, which have the same chemical composition but different crystal structures and, consequently, often different luminescence properties. Despite this known limitation, we deliberately adopted this simplified feature set to prioritize model interpretability, a central goal of this study. It is also noteworthy that sulfur (S) in our dataset can exist as either a cation (S^{6+} in sulfates) or an anion (S^{2-} in sulfides). Because the coordination to Eu^{2+} ions and the effect on Eu 5d-

level energy are expected to differ significantly between these two forms, we treated them as distinct species in this analysis.

In the selection of the feature set, we considered the following points: Previous studies have used statistical measures (e.g. weighted average and standard deviation) of elemental physical properties as features. Although these statistical descriptors are useful for representing diverse macroscopic characteristics of the entire compound, they make it inherently difficult to isolate and evaluate the contribution of each element to the target variable. The primary goal of this study is to maximize model interpretability and quantitatively assess the contribution of each element. To achieve this, we intentionally excluded statistical descriptors and used atomic fractions directly, which represent the most fundamental compositional information of the host materials. This choice is based on a common approach in materials science, where the macroscopic properties of a complex multicomponent system are approximated as a linear sum (additivity rule) of the contributions from its constituent elements. This approach allows the coefficients of the resulting linear model to be directly interpreted as the contribution of each element.

2.3. Model selection and formulation

In this study, we selected ridge regression, a type of linear model, as the machine learning method. We made this choice after a comparative review of other representative linear regression methods with consideration of the objectives of our study and the nature of our features.

Standard linear regression, or ordinary least squares (OLS), maximizes only the goodness of fit to the training data but has the disadvantage that the model can become unstable when features are highly correlated or when the number of data points is small. To address this issue, methods that introduce a regularization term to penalize model complexity are widely used. Representative examples include ridge, LASSO, and elastic-net regression. Ridge regression uses L2 regularization, which penalizes the sum of the squared coefficients. This reduces the overall magnitude of the coefficients, thereby preventing overfitting. In contrast, LASSO regression uses L1 regularization, which penalizes the sum of the absolute values of the coefficients. This property allows LASSO to shrink the coefficients of less important features to exactly zero, thus performing automatic feature selection and creating a simpler (sparse) model. Elastic-net regression combines both L1 and L2 regularization. LASSO and elastic-net are highly effective for improving predictive performance by eliminating features that do not contribute to the prediction. However, the primary goal of this study is not feature selection,

but to comprehensively quantify and interpret the contribution of all constituent elements to the peak emission wavelength, no matter how small. For this purpose, LASSO and elastic-net, which would exclude the contributions of some elements a priori, are unsuitable. In contrast, ridge regression allows all elements to have non-zero contributions while still penalizing their magnitudes. This was the first reason we judged ridge regression to be the most suitable method for our objective.

Second, and more importantly, there is gauge freedom in the coefficients arising from the use of atomic fractions as the explanatory variables, as detailed in the next section. Sparsity-inducing methods, such as LASSO and elastic-net regression, may set coefficients to zero in a specific gauge (reference frame) without consideration of this freedom, leading to coefficients that may not guarantee a unique interpretation. In contrast, ridge regression only penalizes the overall magnitude of the coefficients, leaving room to handle this gauge freedom mathematically and rigorously during the interpretation phase.

In ridge regression, for a host compound i with an emission wavelength $\lambda_{\text{em},i}$, its predicted value $\hat{\lambda}_{\text{em},i}$ is given by the following equation:

$$\hat{\lambda}_{\text{em},i} = \beta_0 + \sum_j \beta_j x_{ij}$$

where x_{ij} is the atomic fraction of element j in host i . The intercept β_0 and the coefficients β_j are the parameters to be determined by minimizing the loss function L , which includes an L2 regularization term, as shown below:

$$L = \sum_i \left(\hat{\lambda}_{\text{em},i} - \lambda_{\text{em},i} \right)^2 + \alpha \sum_{j \neq 0} \beta_j^2$$

Here, the hyperparameter α represents the strength of the regularization and was optimized to minimize the average root mean squared error (RMSE) in a 10-fold cross-validation (CV). The details of hyperparameter determination are described in Section S1 of the Supplemental material. The model construction, CV, and bootstrap sampling described in Section 2.5 were performed using the scikit-learn ML library (version 1.2.2) [19].

2.4. Definition of the elemental contribution coefficient (ECC)

The coefficients β_j obtained from ridge regression reflect the influence of each element on the predicted emission wavelength, but their raw values cannot be directly interpreted in a straightforward manner. This is because the atomic fractions x_{ij} , used as features in this study, must satisfy two linear constraint conditions: (1)

$\sum_j x_{ij} = 1$ (the sum of fractions is 1) and (2) $\sum_j q_j x_{ij} = 0$ (charge neutrality, where q_j is the oxidation number of element j). These constraints make it possible to transform the intercept β_0 and coefficients β_j into the form $\beta'_0 = \beta_0 - a$ and $\beta'_j = \beta_j + a + bq_j$, where a and b are arbitrary constants, without changing predictions by the model $\hat{\lambda}_{\text{em},i}$. This indeterminacy, known as gauge freedom, particularly the bq_j term, introduces a shift to the coefficients that depends on the oxidation numbers q_j , making a direct comparison of coefficients between elements with different oxidation numbers principally impossible.

To resolve this gauge freedom and obtain a consistent and comparable representation of elemental contributions, it is necessary to define a new metric. In this study, we introduced an elemental contribution coefficient (ECC) as an analytical framework to address this challenge. The ECC aims to enable a mutual comparison of the contributions of different elements under the assumptions of the model, independent of the specific host composition or gauge choice. Its derivation follows these steps:

First, to cancel the common offset term arising from the arbitrary constant a , we introduce a reference contribution $\bar{\beta}$ and evaluate the contribution of each element as the difference $\beta_j - \bar{\beta}$. Although the definition of $\bar{\beta}$ is arbitrary, we adopt the most natural choice: the weighted average of the coefficients for the elements present in the dataset, $\bar{\beta} = \sum_j \beta_j \bar{x}_j$. Here, the weights $\bar{x}_j$ are set to a common value $\bar{x}_{\text{cation}}$ for all cations and $\bar{x}_{\text{anion}}$ for all anions, such that $\sum_j \bar{x}_j = N_{\text{cation}} \bar{x}_{\text{cation}} + N_{\text{anion}} \bar{x}_{\text{anion}} = 1$, where N_{cation} and N_{anion} are the number of cation and anion types, respectively. For $\bar{\beta}$ to serve as a gauge-invariant reference point, unaffected by the gauge transformation b , it must satisfy the condition $\sum_j q_j \bar{x}_j = \left(\sum_{j \in \text{cations}} q_j \right) \bar{x}_{\text{cation}} + \left(\sum_{j \in \text{anions}} q_j \right) \bar{x}_{\text{anion}} = 0$. These conditions uniquely determine the weights $\bar{x}_{\text{cation}}$ and $\bar{x}_{\text{anion}}$.

Second, to isolate the gauge freedom term bq_j , which is proportional to the charge and remains in the difference $\beta_j - \bar{\beta}$, we divide this difference by the absolute value of the oxidation number of each element $|q_j|$. We define this as the elemental contribution coefficient (ECC), γ_j :

$$\gamma_j = \frac{\beta_j - \bar{\beta}}{|q_j|}$$

Through this operation, the arbitrary gauge constant b is separated as a common additive offset term for all cations (or, separately, for all anions). Finally, this offset term is removed by shifting the γ_j values so that the average for all cations (and separately for

all anions) becomes zero. This yields a gauge-invariant, mutually comparable elemental contribution metric.

The resulting ECC, γ_j , can be interpreted as a model-derived index representing the relative tendency of element j to influence the predicted emission wavelength per unit charge. It should be noted that the ECC is not an intrinsic or universal physical quantity. Rather, it is a dataset-conditioned index defined within the context of the present model and dataset. Therefore, the ECC should be interpreted as a relative measure of elemental tendencies under the assumptions of the model, rather than as a transferable material descriptor. With this reservation in mind, we will use the derived ECCs to discuss the trends associated with each element.

2.5. Uncertainty quantification of the ECC

The reliability of the ECC, the dataset-conditioned index derived in this study, depends on the statistical significance of the calculated coefficients. A common challenge with real-world datasets, which our dataset also shares, is an imbalanced distribution (bias) of the number of compounds containing each constituent element. Some elements appear frequently, while others appear only rarely, as described in Section 3.1. This imbalance can affect the stability and reliability of the coefficients of the linear model, β_j , and, consequently, the derived ECCs, γ_j .

To quantitatively evaluate this effect, we used the bootstrap method to estimate the confidence intervals for these coefficients. The bootstrap procedure was as follows. First, we generated 10,000 datasets of the same size as the original by resampling from it with replacement. Next, for each resampled dataset, we trained a ridge regression model and calculated the coefficients β_j and the ECCs γ_j . The regularization parameter α was re-optimized for each bootstrap sample using 10-fold CV to account for variations in data distribution. This procedure yields a statistical distribution for the ECCs.

Two points are noteworthy regarding this procedure. First, if a resampled dataset does not contain a particular element, the corresponding coefficient β_j is set to zero in ridge regression. However, this value is meaningless due to the gauge freedom. Therefore, in such cases, the ECC γ_j for that element was treated as a missing value. Second, the bootstrap analysis in this study is intended to evaluate the stability of the ECC with respect to data resampling within the dataset, rather than to provide a measure of predictive uncertainty or absolute physical accuracy.

3. Results

3.1. Dataset characteristics

The basic characteristics of the dataset analyzed in this study are described below. Figure 1 shows the distribution of the peak emission wavelengths, $\lambda_{em,i}$, for the 118

Eu^{2+} -activated phosphors in the dataset. The emission wavelengths range from 375 nm to 734 nm, covering the entire visible spectrum. The distribution exhibits two prominent peaks, indicating that the data are concentrated in the blue (400–440 nm) and green (500–540 nm) regions.

Figure 2 shows the number of compounds in the dataset containing each constituent element. Among the anions, O^{2-} and N^{3-} are the most common. Among the cations, Ca^{2+} , Sr^{2+} , and Ba^{2+} , which are selected as substitution sites for Eu^{2+} , and Mg^{2+} , Al^{3+} , Si^{4+} , and P^{5+} , which form the framework of the crystal structure, appear in many compounds. This combination of elements indicates that the dataset is primarily composed of oxides, nitrides, and their complex compounds (e.g. silicates and phosphates). This is a result of the data collection process, which reflects recent research trends in the pc-WLED applications. Consequently, a sampling

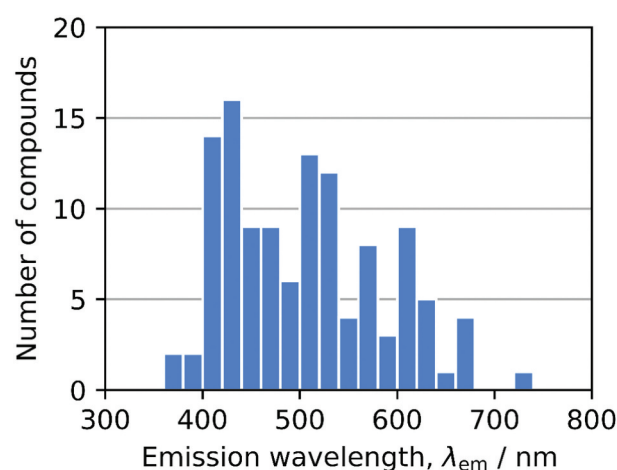


Figure 1. Distribution of the peak emission wavelengths of the 118 Eu^{2+} -activated phosphors in the curated dataset.

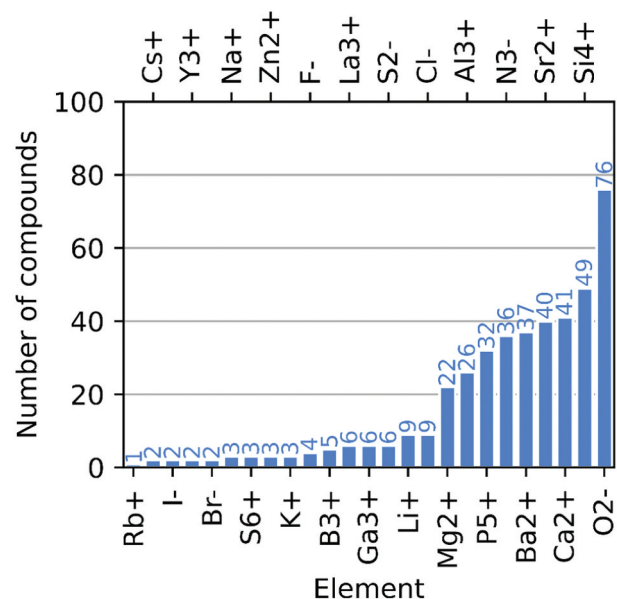


Figure 2. Number of compounds in the dataset containing each constituent element.

bias is present, with data on classical phosphor materials such as halides and sulfides in the minority.

Furthermore, the distribution of element occurrences exhibits a typical long-tail shape: A few elements are found in many compounds, whereas many elements appear in only a few compounds. These characteristics of the dataset, namely, the chemical species bias and this imbalanced distribution, may affect the performance of the model constructed in this study and the interpretation of the insights obtained. In particular, careful evaluation is required to assess the reliability of the model for elements found in only a very small number of compounds.

3.2. Predictive performance

The predictive performance of the ridge regression model was evaluated using 10-fold CV. Figure 3 shows a plot of the predicted versus reported values from the CV at the optimal hyperparameter $\alpha = 0.01$. The data points in both the training and validation sets are approximately distributed along the diagonal line. Table 1 summarizes the average and standard deviation of the mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) from the CV. The average MAE and RMSE on the validation data were 33.4 nm and 40.5 nm, respectively. Considering that the peak emission wavelength of phosphors can vary by about 10 nm depending on experimental conditions, such as Eu^{2+} concentration, and that the full width at half maximum (FWHM) of an emission spectrum of a typical Eu^{2+} -activated phosphor is 50–100 nm, this level of error

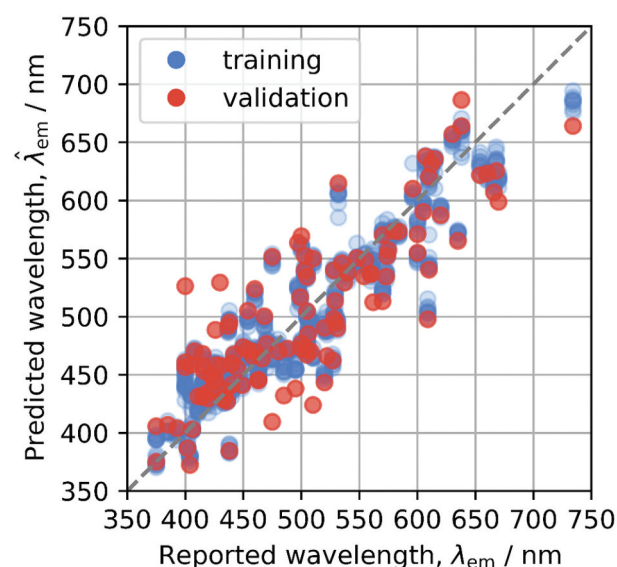


Figure 3. Predicted versus reported emission wavelengths from the 10-fold cross-validation. The blue and red circles represent the training and validation data for each fold, respectively.

Table 1. Performance metrics of the prediction model, including mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). The values are averaged over the 10 folds of cross-validation. The values in parentheses represent the standard deviations across the folds.

Metrics	Training	Validation
MAE (nm)	25.6 (1.1)	33.4 (8.3)
RMSE (nm)	32.4 (1.1)	40.5 (10.7)
R^2	0.83 (0.01)	0.70 (0.15)

is practical for initial screening in the search for new host materials.

Furthermore, the plot in Figure 3 shows no significant discrepancy between the distributions of predicted values for the training and validation data for most compounds, suggesting that our model does not suffer from severe overfitting. However, some compounds exhibit large prediction errors. The presence of these outliers may be attributed to the simplicity of the model and features adopted in this study, or to other factors, such as data quality or differences in the luminescence mechanisms of certain compounds.

It should be noted that, because we curated the dataset from the previous study [10], our final dataset is not identical to it. Therefore, a direct comparison of predictive accuracy with the previous study is not possible. For reference, we applied the same procedure as in this study to the dataset from the previous study, and the validation MAE was 30.6 (5.4) nm, RMSE was 38.5 (5.9) nm, and R^2 was 0.73 (0.10). These metrics are slightly lower than but comparable to those of the ridge regression model from the previous study, which used more complex features (MAE = 29 nm, RMSE = 36 nm, $R^2 = 0.74$).

3.3. Elemental contribution coefficient (ECC)

Figure 4 shows the calculated ECCs, plotted separately for anions and cations. Among the anions, N^{3-} shows the largest positive value (a redshift contribution), whereas Br^- shows the largest negative value (a blueshift contribution). The other halide ions (F^- , Cl^- , and I^-) also have large negative values, whereas O^{2-} and S^{2-} have intermediate values between them.

Among the cations, alkali metal ions such as Na^+ and alkaline earth metal ions such as Ca^{2+} tend to show large positive values. In contrast, high-valence cations such as S^{6+} and P^{5+} exhibit large negative values. The trends observed in these ECC values are discussed in detail in Section 4. For reference, the raw coefficients β_j of the linear model are shown in Figure S2 in the Supplemental material.

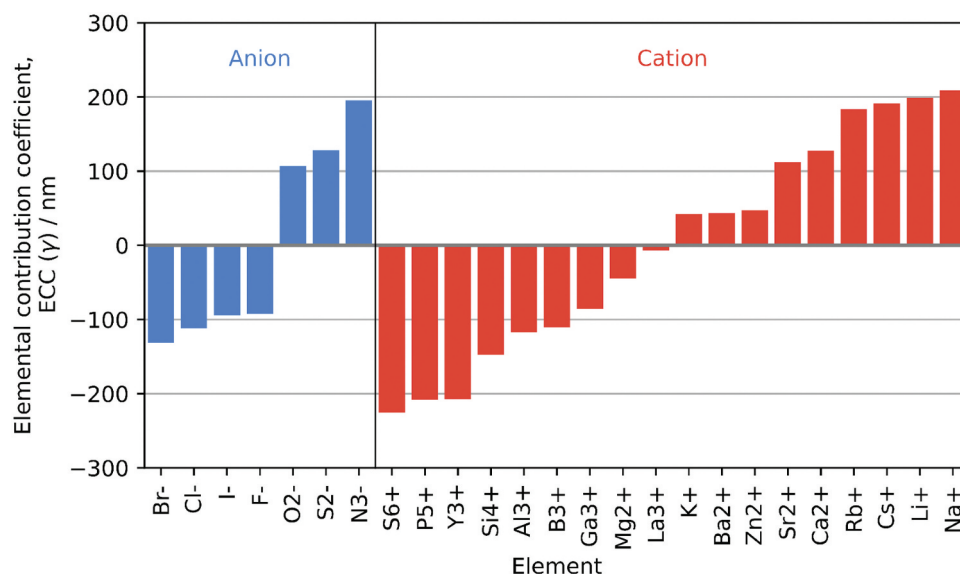


Figure 4. Elemental contribution coefficient (ECC), γ_j , for each element, derived from the coefficients of the linear model. The ECCs are plotted separately for anions and cations.

3.4. Uncertainty quantification of the ECC

To assess the statistical reliability of the calculated ECCs, Figure 5 shows box plots of the ECC distributions obtained using the bootstrap method. The elements on the horizontal axis are arranged in the same order as the ECCs calculated from the full dataset (Figure 4). The order of the median values of the ECC distributions (the line inside each box) is generally consistent with the order shown in Figure 4.

On the other hand, the confidence intervals of the ECCs, evaluated here using the interquartile range (IQR), vary significantly among the elements. In particular, the IQRs for alkali metal ions K^+ , Rb^+ , and Cs^+ are notably wider than those for other element groups.

Two factors are considered to be the cause of this phenomenon. First, these elements appear infrequently, occurring in at most three compounds in the dataset. The second is the error amplification effect arising from the definition of the ECC ($\gamma_j = (\beta_j - \bar{\beta}) / |q_j|$). For monovalent cations, where the absolute charge $|q_j|$ is 1, any statistical uncertainty in the coefficient β_j is directly reflected in the uncertainty of the ECC. In contrast, for high-valence cations, the error is relatively compressed because $|q_j|$ is large.

However, the width of the confidence interval is not solely determined by the amount of data. For example, Br^- and I^- showed relatively narrow IQRs despite their

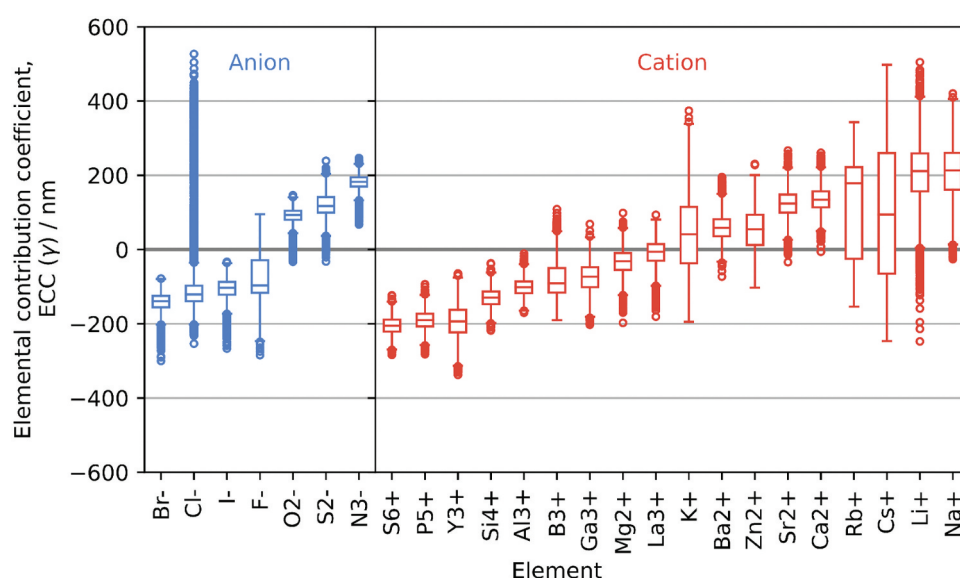


Figure 5. Box plots of the elemental contribution coefficient (ECC) distributions obtained from 10,000 bootstrap samples. The arrangement of the elements on the x-axis is the same as that in Figure 4. Each box represents the interquartile range (IQR) from the 25th to 75th percentiles, with the central line indicating the median.

low occurrence and monovalence. This suggests that factors beyond the data count influence uncertainty. As another characteristic example, Cl^- has a narrow IQR, but its distribution shows a long tail extending towards the outlier direction. These observations suggest that the uncertainty of the calculated ECCs depends not only on the amount of data but also on the diversity of the chemical environments (data distribution) for each element within the dataset. A detailed discussion on the distribution of Cl^- is provided in Section 4.2.1. These results indicate that when interpreting the contribution of a specific element, its statistical confidence interval must be considered individually.

4. Discussion

This study adopted an approach that prioritizes interpretability over predictive accuracy to address the common interpretability dilemma in materials informatics. We constructed a simple linear model using only atomic fractions as features to quantify the contribution of each element to the emission wavelength, yielding an interpretable coefficient within the model. In this section, we discuss both the insights obtained from the analysis and the limitations inherent to the modeling framework. First, we evaluate the qualitative consistency of the model by examining how well it reproduces the established empirical rules in phosphor chemistry (Section 4.1). Second, by analyzing the discrepancies between our results and these rules, we clarify the fundamental limitations of our approach and identify key physical factors for future consideration (Section 4.2).

4.1. Agreement with empirical rules

As shown in Section 3.2, our linear model achieved a practical level of predictive accuracy using only the overall atomic fractions of the host compound as input. This fact raises the question: Why can the emission wavelength be predicted to some extent without explicitly considering the local structural information around the luminescence center? This is likely because the overall average composition (macroscopic information) imposes crystallochemical constraints on the local coordination environment around the luminescence center (microscopic information). For instance, if a host consists of a single anion species, the macroscopic composition directly determines the microscopic coordination anion. Furthermore, stable crystal structures are governed by the balance of size and charge of the constituent ions, as represented by Pauling's rules [20]. For example, the first rule (the radius ratio rule) determines the coordination number based on the ionic size ratio, and the second rule (the electrostatic valence rule)

constrains the local bonding arrangement based on ionic charges. Therefore, the macroscopic average composition indirectly limits the range of permissible local coordination structures. However, this constraint is not always strict enough to define a single structure, often allowing for the existence of polymorphs, different crystal structures with the same chemical composition. Our model, which relies solely on atomic fractions, cannot, in principle, distinguish between these polymorphs. Nevertheless, the observation that a given composition statistically favors a particular range of local environments implies that information about the microscopic structure is embedded in the macroscopic composition. This supports the predictive ability of our atomic fraction model, despite this inherent limitation.

To further examine the consistency of this model, we compared trends in the calculated ECCs, γ_j , with empirical physical rules. The results showed a qualitative agreement between them. First, looking at the contribution of the second-period anions, the ECC increases in the order $\text{F}^- < \text{O}^{2-} < \text{N}^{3-}$, with N^{3-} showing the strongest redshift contribution. This corresponds to the effect of the electron affinity of the anion coordinating on the luminescence center, as noted by Van Uitert. This trend is also consistent with the nephelauxetic effect, as systematized by Dorenbos, in which the center of gravity of the 5d-level energy decreases as the covalency of the anion increases. The same trend is observed among the third-period anions: $\text{Cl}^- < \text{S}^{2-}$.

Next, for the alkaline-earth-metal sites substituted by Eu^{2+} , the ECC increases in the order $\text{Ba}^{2+} < \text{Sr}^{2+} < \text{Ca}^{2+}$. This is consistent with the empirical rules of Van Uitert and Dorenbos, which state that sites with smaller ionic radii exhibit greater crystal field splitting, leading to a shift to longer wavelengths in the emission spectrum. To further examine the consistency of the ECC with compositional trends, we analyzed pairs of host compounds in which Ca, Sr, or Ba are substituted within similar compositional frameworks. 23 Ca-Sr pairs and 19 Sr-Ba pairs were identified in the dataset, excluding $\text{Ca}_3(\text{PO}_4)_2$, which has polymorphs. For each pair, the difference in peak emission wavelength relative to the Sr-based host was evaluated. Figure 6 shows the resulting distributions, indicating that Ca-containing hosts tend to exhibit longer emission wavelengths than the corresponding Sr-containing hosts, whereas Ba-containing hosts tend to show shorter emission wavelengths. These results are generally consistent with the empirical trend ($\text{Ca} > \text{Sr} > \text{Ba}$). However, the distributions exhibit large variances, and the sign of the shift is not uniform across all cases. This indicates that while the overall tendency is consistent with the ECC trends, individual systems may deviate significantly due to additional factors not captured by the composition-based model. These observations support the

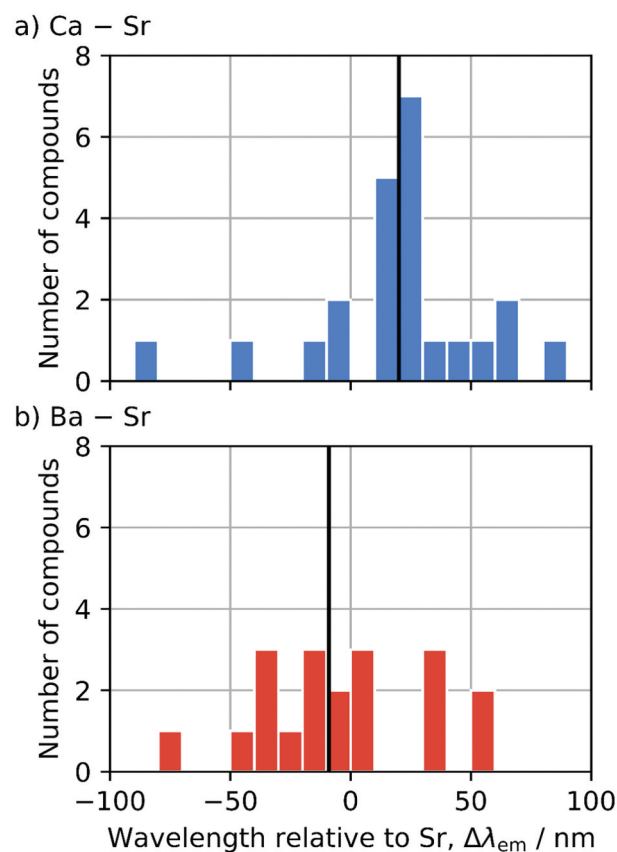


Figure 6. Distribution of differences in peak emission wavelength relative to Sr-based hosts for (a) Ca-containing and (b) Ba-containing compounds in paired systems with similar compositional frameworks. Black lines indicate medians of the differences among the Ca and Ba-containing systems. The differences are defined as $\Delta\lambda_{em} = \lambda_{em}(\text{Ca or Ba}) - \lambda_{em}(\text{Sr})$ for each pair. Positive values indicate a redshift relative to Sr, while negative values indicate a blueshift.

interpretation that the ECC reflects dataset-level compositional tendencies rather than deterministic predictions for individual compounds. It is noteworthy that the trends observed in this subset analysis do not necessarily coincide exactly with those inferred from the ECC, as the latter reflects regression over the entire dataset and is therefore influenced by a broader distribution of compositions and chemical environments.

Furthermore, even for the cations that form the host crystal framework, the trend in ECCs (e.g. $\text{P}^{5+} < \text{Si}^{4+} < \text{Al}^{3+}$) generally agrees with the trend reported by Dorenbos, where more electronegative cations suppress the polarization of anions, resulting in a smaller centroid shift of the 5d levels (i.e. a blueshift contribution). Unfortunately, extending the analysis of relative peak emission wavelengths to other elements, particularly those with different valences, is more challenging because of differences in oxidation states and structural roles, unlike the isovalent substitutions of alkaline-earth elements (Ca, Sr, and Ba). This complicates direct comparison within the present dataset. However, the trends observed in the ECC are consistent with the common

empirical rules in phosphor chemistry. These observations support the use of the ECC as a tool for extracting trends in the dataset.

4.2. Model limitations and their implications

In this section, we discuss both the limitations revealed through the analysis and those inherent to the modeling framework.

4.2.1. Limitations revealed by the analysis

As discussed in Section 4.1, there is a strong agreement in many respects between the calculated ECCs and empirical rules. On the other hand, several systematic discrepancies were also observed. These discrepancies reveal the fundamental limitations of our linear, composition-based approach adopted in this study: It cannot fully describe physical effects that are not solely determined by composition.

The most prominent example is the trend among anions. According to the nephelauxetic series presented by Dorenbos, the ECC is expected to increase in the order $\text{O}^{2-} < \text{Cl}^-$, Br^- , and I^- based on covalency. However, as shown in Figure 4, the ECC values calculated by our model for halide ions (Cl^- , Br^- , and I^-) are smaller (with more blueshift contributions) than that of O^{2-} , seemingly contradicting the expectation based on the covalency trend. This apparent contradiction can be interpreted by recognizing that our model simultaneously learns multiple effects that sometimes act in opposite directions, and outputs their ‘net effect’. According to Dorenbos, the emission wavelength shift (redshift) is determined by the sum of two independent physical effects: the centroid shift, which primarily depends on covalency (a chemical effect), and crystal field splitting, which primarily depends on the geometric arrangement. For instance, let us compare Cl^- , which has a larger ionic radius, with O^{2-} , which is smaller. Cl^- causes a larger centroid shift due to its stronger nephelauxetic effect (a redshift contribution), but its larger ionic radius leads to a smaller crystal field splitting (a blueshift contribution). The fact that the ECC of Cl^- is smaller than that of O^{2-} in our model can be interpreted as the latter blueshift effect outweighing the former redshift effect. This suggests that the data contain complexities of physical phenomena that cannot be captured by simple intuition. Our linear model reflects this complexity, but it cannot separate the underlying competing effects. A similar argument can be applied to the trend between N^{3-} and S^{2-} .

In addition to the entanglement of physical effects discussed above, the inherent bias in the dataset must also be considered when interpreting learning outcomes of the model. As shown in Figure 2, the dataset is dominated by oxides and nitrides, with halides and sulfides as minorities. Therefore, the constructed

model has primarily learned chemical trends in oxides and nitrides. Careful consideration is needed to determine whether we can apply the learned rules equally to minority compound groups. For example, the halides in this dataset are largely limited to compositionally simple compounds, such as CaBr_2 , SrI_2 , and RbCaF_3 . In contrast, the oxides and nitrides, which constitute the majority of the dataset, form more diverse and complex network structures, such as those found in silicates and phosphates. Such systematic differences in compositional and structural complexity may be learned by the model as a secondary effect associated with the anions, making it inseparable from the pure effect of the anion (e.g. the nephelauxetic effect). This could be another factor that complicates the straightforward comparison between ECC trends and empirical rules.

We can perform a similar analysis of trends among the alkaline earth metals. The ECCs for alkaline earth metals, whether calculated from the entire dataset (Figure 4) or the median values (Figure 5), show an increasing trend in the order $\text{Ba}^{2+} < \text{Sr}^{2+} < \text{Ca}^{2+}$. This is consistent with the empirical rule that a smaller ionic radius leads to a larger crystal field splitting. However, the bootstrap evaluation results in Figure 5 suggest a more complex situation. The confidence intervals for the ECCs of Sr^{2+} and Ca^{2+} largely overlap, indicating that the difference between them is not robust with respect to data resampling within the present dataset. It should be noted that these confidence intervals reflect the stability of the ECC under resampling, rather than the uncertainty of an underlying physical quantity. This suggests that the relative contributions of Sr^{2+} and Ca^{2+} are not consistently distinguished by the present dataset and model. One possible interpretation is that multiple factors may influence the observed trends. For example, the ionic radius ($\text{Ba}^{2+} > \text{Sr}^{2+} > \text{Ca}^{2+}$) and electronegativity ($\text{Ba}^{2+} < \text{Sr}^{2+} < \text{Ca}^{2+}$) of alkaline earth elements are inversely correlated. The former increases crystal field splitting (a redshift for Ca^{2+}). In contrast, the latter, according to Dorenbos's rule, leads to a suppression of the centroid shift (a blueshift contribution for Ca^{2+}). In other words, the ECCs for alkaline earth metals, similar to the case of anions, can be understood as reflecting a mixture of competing effects that act in opposite directions. The lack of a clear difference between the ECCs of Sr^{2+} and Ca^{2+} may be interpreted as the result of these competing effects canceling each other to a similar degree for both elements, yielding a net effect that is similar for both.

The preceding discussions of the anion and cation trends clarify a common fundamental limitation of our approach: Compositional information alone cannot distinguish between competing physical effects, such as the centroid shift and crystal field splitting.

This dependence on compositional information becomes particularly apparent when the implicit correlation between macroscopic composition and the microscopic local environment breaks down. A clear example of this is the low statistical reliability of the ECC for Cl^- (see Figure 5), as reflected in its sensitivity to data resampling. The long tail observed in the ECC distribution for Cl^- is caused by the inclusion of both compositionally simple compounds, such as CaCl_2 , and complex mixed-anion compounds, such as $\text{Ca}_5(\text{PO}_4)_3\text{Cl}$, where the implicit correlation between macroscopic composition and microscopic local environment breaks down. In $\text{Ca}_5(\text{PO}_4)_3\text{Cl}$, despite the low overall atomic fraction of Cl^- , one of the two Ca sites (Ca1) is 7-coordinated, with two of its coordinating anions being Cl^- . This local environment differs significantly from the average overall composition. Our model, which uses only atomic fractions as features, cannot, in principle, handle such diversity in local environments, leading to instability in the calculation of ECC for Cl^- within the present dataset and modeling framework.

These considerations extend beyond merely highlighting the limitations of the current model. This raises the following scientific question: What additional information is needed to separate these competing effects and capture the diversity of local environments? This strongly suggests that a critical challenge for future research in phosphor informatics is not only the search for features to improve predictive accuracy but also the development of more interpretable features that can separate and interpret physical effects. To achieve both ultimate predictive accuracy and interpretability, it is essential to develop more sophisticated features that can represent the local coordination environment around the luminescence center with interpretability comparable to that of atomic fractions.

4.2.2. Methodological limitations of the framework

In addition to the limitations discussed in Section 4.2.1, our approach has other limitations inherent to the modeling framework. First, the present model relies solely on atomic fractions and a linear formulation. Therefore, the model cannot explicitly account for interaction effects between elements. More expressive models, such as tree-based methods combined with interpretability techniques (e.g. SHAP), may capture nonlinear interactions between elements. However, their interpretation in terms of chemical intuition is not always straightforward, especially when complex feature interactions are involved. Developing interpretable representations that can incorporate such effects while maintaining transparency remains an important challenge.

Second, the treatment of sulfur as separate species (S^{6+} and S^{2-}) is somewhat ad hoc. More systematic

representations of chemical states for elements with multiple oxidation states would improve the consistency of the framework.

Third, the dataset itself introduces additional limitations, including the exclusion of activator concentration effects and potential sampling bias. In practice, Eu^{2+} concentration can significantly influence emission wavelengths. However, it is difficult to obtain consistent and reliable concentration-dependent data from the literature due to variations in synthesis conditions and reporting standards. In this study, we deliberately restricted the model to host compositions to extract robust compositional trends. Incorporating activator concentration as an additional feature would further enhance the applicability of the model.

Finally, because the ECC is defined based on dataset-dependent coefficients and normalization, its values may change as the dataset evolves. This reflects the fact that the ECC captures trends embedded in the available data rather than intrinsic, universally defined properties of elements. Therefore, the ECC should be interpreted within the scope of the dataset and modeling assumptions used in this study. While the values of the ECC are expected to depend on specific data curation criteria, measurement conditions, and chemical space coverage, the qualitative trends may be more robust if the dataset is sufficiently diverse and high-quality. This is supported by the observed consistency of the extracted trends with established chemical intuition. Accordingly, the present analytical framework can, in principle, be applied to independent datasets, provided that basic consistency in data treatment is maintained. However, direct comparison or combination of ECC values obtained from different datasets may not be appropriate without careful alignment, as systematic differences in data characteristics can lead to shifts in the resulting values. Establishing robust cross-dataset comparability, therefore, remains an important challenge.

These limitations should be considered when interpreting the ECC as a trend-analysis tool. From a practical perspective, the present framework is particularly suited to early-stage screening and trend analysis, where rapid, interpretable insights into compositional tendencies are more valuable than precise quantitative predictions. At the same time, the approach is not limited to this use case, and its primary contribution is to provide a transparent analytical framework for understanding trends in the dataset.

5. Conclusions

This study has demonstrated a new path for data-driven science to extract interpretable trends in materials by prioritizing interpretability to address the fundamental trade-off between predictive accuracy and interpretability

in modern materials informatics. The core contribution of our work is the demonstration that a simple, interpretable model based on atomic fractions can serve as a useful tool for analyzing compositional trends in phosphor materials. This tool provides a way to examine ‘why’ trends are observed in the data, in addition to ‘what’ is predicted, through the elemental contribution coefficient (ECC). Specifically, we have revealed two capabilities of this tool. The first is the ability to recover trends consistent with chemical intuition and to interpret data in terms of interpretable trends. We showed that the calculated ECCs qualitatively agree with classical empirical rules, such as the nephelauxetic effect and crystal field splitting. The second is the ability to detect deviations from empirical rules and raise new questions. By closely examining the apparent contradictions between the ECCs and empirical rules, we identified their cause as the entanglement of competing physical effects, such as the centroid shift and crystal field splitting, which our linear, compositional-based approach cannot separate. The proposed ECC provides a transparent, data-driven index that complements chemical intuition by making implicit trends in the dataset explicit.

Of course, this approach has its fundamental limitations. It cannot distinguish between polymorphs with the same chemical composition, and its predictive reliability decreases in systems where the correlation between the macroscopic composition and the microscopic local structure breaks down, as in mixed-anion compounds. Rather than serving as a universal or physically rigorous descriptor, the ECC is best understood as a framework for trend analysis and hypothesis generation within a given dataset. The limitations identified in this study, including the inability to capture local structural effects and competing physical mechanisms, point to important directions for future development of interpretable features in materials informatics: the development of next-generation interpretable local-structural features. These features should be able to separate and capture competing physical effects, such as the centroid shift and crystal field splitting, while maintaining a high level of interpretability compared to atomic fractions. Such an approach would be most promising for its effectiveness in material systems governed by similar physical mechanisms, such as Ce^{3+} -activated phosphors that are also based on the 4f-5d transition. Furthermore, this approach is applicable to many other inorganic functional material systems, including transition-metal compounds, in which local environments strongly govern properties. It may provide a useful basis for developing interpretable approaches to materials design across related material systems. We are confident that if realized, this will establish the way for high-precision, interpretable approaches to materials design.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Yukinori Koyama  <http://orcid.org/0000-0002-7090-4430>

Yuta Matsushima  <http://orcid.org/0000-0001-5826-1551>

Takayuki Nakanishi  <http://orcid.org/0000-0003-3412-2842>

Takashi Takeda  <http://orcid.org/0000-0003-2510-4562>

Data availability statement

The dataset used in this study is available as a comma-separated values (CSV) file in the Supplemental material.

References

- [1] Ye S, Xiao F, Pan YX, et al. Phosphors in phosphor-converted white light-emitting diodes recent advances in materials, techniques and properties. *Mat Sci Eng R*. 2010;71(1):1–34. doi: [10.1016/j.mser.2010.07.001](https://doi.org/10.1016/j.mser.2010.07.001)
- [2] Xia ZG, Liu QL. Progress in discovery and structural design of color conversion phosphors for LEDs. *Prog Mater Sci*. 2016;84:59–117. doi: [10.1016/j.pmatsci.2016.09.007](https://doi.org/10.1016/j.pmatsci.2016.09.007)
- [3] Qin X, Liu XW, Huang W, et al. Lanthanide-activated phosphors based on 4f-5d optical transitions: theoretical and experimental aspects. *Chem Rev*. 2017;117(5):4488–4527. doi: [10.1021/acs.chemrev.6b00691](https://doi.org/10.1021/acs.chemrev.6b00691)
- [4] Wang L, Xie RJ, Suehiro T, et al. Down-conversion nitride materials for solid state lighting: recent advances and perspectives. *Chem Rev*. 2018;118(4):1951–2009. doi: [10.1021/acs.chemrev.7b00284](https://doi.org/10.1021/acs.chemrev.7b00284)
- [5] Van Uitert LG. An empirical relation fitting the position in energy of the lower d-band edge for Eu-2+ or Ce-3+ in various compounds. *J Lumines*. 1984;29(1):1–9. doi: [10.1016/S0022-2313\(84\)90070-X](https://doi.org/10.1016/S0022-2313(84)90070-X)
- [6] Dorenbos P. A review on how lanthanide impurity levels change with chemistry and structure of inorganic compounds. *ECS J Solid State Sci Technol*. 2013;2(2):R3001–R3011. doi: [10.1149/2.001302jss](https://doi.org/10.1149/2.001302jss)
- [7] Park WB, Singh SP, Kim M, et al. Phosphor informatics based on confirmatory factor analysis. *ACS Comb Sci*. 2015;17(5):317–325. doi: [10.1021/acs.combsci.5b00017](https://doi.org/10.1021/acs.combsci.5b00017)
- [8] Park C, Lee JW, Kim M, et al. A data-driven approach to predicting band gap, excitation, and emission energies for Eu2+-activated phosphors. *Inorg Chem Front*. 2021;8(21):4610–4624. doi: [10.1039/d1qi00766a](https://doi.org/10.1039/d1qi00766a)
- [9] Nakano H, Tanaka K, Miyao T, et al. Practical models for predicting the emission peak wavelengths of inorganic phosphors based on stoichiometric information. *Chem Lett*. 2017;46(10):1482–1485. doi: [10.1246/cl.170611](https://doi.org/10.1246/cl.170611)
- [10] Koyama Y, Ikeno H, Harada M, et al. Rapid discovery of new Eu2+-activated phosphors with a designed luminescence color using a data-driven approach. *Mater Adv*. 2023;4(1):231–239. doi: [10.1039/D2MA00881E](https://doi.org/10.1039/D2MA00881E)
- [11] Zhang Z, Zhou J, Wen J, et al. Machine learning analysis and discovery of novel blue-cyan-emitting Na18Ca13Mg5(PO4)18: eu2+ phosphor. *J Lumines*. 2025;286:121408. doi: [10.1016/j.jlumin.2025.121408](https://doi.org/10.1016/j.jlumin.2025.121408)
- [12] Shi X, Zhong X, Liu W, et al. Prediction and exploration of emission wavelength (or energy) of luminescent materials based on machine learning. *J Lumines*. 2025;279:121024. doi: [10.1016/j.jlumin.2024.121024](https://doi.org/10.1016/j.jlumin.2024.121024)
- [13] Jang S, Na GS, Choi Y, et al. Optical property dataset of inorganic phosphor. *Sci Rep*. 2024;14(1):10. doi: [10.1038/s41598-024-58351-w](https://doi.org/10.1038/s41598-024-58351-w)
- [14] Dean J, Scheffler M, Purcell T, et al. Interpretable machine learning for materials design. *J Mater Res*. 2023;38(20):4477–4496. doi: [10.1557/s43578-023-01164-w](https://doi.org/10.1557/s43578-023-01164-w)
- [15] Sivan D, Kumar K, Abdullah A, et al. Advances in materials informatics: a review. *J Mater Sci*. 2024;59(7):2602–2643. doi: [10.1007/s10853-024-09379-w](https://doi.org/10.1007/s10853-024-09379-w)
- [16] Jiang X, Fu H, Bai Y, et al. Interpretable machine learning applications: a promising prospect of AI for materials. *Adv Funct Mater*. 2025;35(41):2507734. doi: [10.1002/adfm.202507734](https://doi.org/10.1002/adfm.202507734)
- [17] Otyepka M, Pykal M, Otyepka M. Advancing materials discovery through artificial intelligence. *Appl Mater Today*. 2025;47:102981. doi: [10.1016/j.apmt.2025.102981](https://doi.org/10.1016/j.apmt.2025.102981)
- [18] Yen WM, Weber MJ. Inorganic phosphors: compositions, preparation and optical properties. Boca Raton: CRC Press; 2004.
- [19] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
- [20] Pauling L. The principles determining the structure of complex ionic crystals. *J Am Chem Soc*. 1929;51(1–4):1010–1026. doi: [10.1021/ja01379a006](https://doi.org/10.1021/ja01379a006)