

Delft University of Technology

### Machine learning 5d-level centroid shift of Ce<sup>3+</sup>inorganic phosphors

Zhuo, Ya; Hariyani, Shruti; You, Shihai; Dorenbos, Pieter; Brgoch, Jakoah

DOI 10.1063/5.0012434

Publication date 2020 Document Version Accepted author manuscript

Published in Journal of Applied Physics

### Citation (APA)

Zhuo<sub>3</sub><sup>+</sup>, Hariyani, S., You, S., Dorenbos, P., & Brgoch, J. (2020). Machine learning 5d-level centroid shift of Ce<sup>-</sup>inorganic phosphors. *Journal of Applied Physics*, *128*(1), Article 013104. https://doi.org/10.1063/5.0012434

### Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

### Machine learning 5d-level centroid shift of Ce<sup>3+</sup> inorganic phosphors

Ya Zhuo<sup>1</sup>, Shruti Hariyani<sup>1</sup>, Shihai You<sup>1,2</sup>, Pieter Dorenbos<sup>3</sup>, and Jakoah Brgoch<sup>1,a)</sup>

### **AFFILIATIONS**

<sup>1</sup>Department of Chemistry, University of Houston, Houston, Texas 77204, USA <sup>2</sup>College of Materials, Xiamen University, Xiamen, Fujian 361005, China <sup>3</sup>Interfaculty Reactor Institute, Delft University of Technology, Mekelweg 15, 2629 JB Delft, The Netherlands <sup>a)</sup>Author to whom correspondence should be addressed: jbrgoch@uh.edu

### ABSTRACT

Information on the 5d level centroid shift ( $\varepsilon_c$ ) of rare-earth ions is critical for determining the chemical shift and the Coulomb repulsion parameter as well as predicting the luminescence and thermal response of rare-earth substituted inorganic phosphors. The magnitude of  $\varepsilon_c$  depends on the binding strength between the rare-earth ion and its coordinating ligands, which is difficult to quantify a *priori* and makes phosphor design particularly challenging. In this work, a tree-based ensemble learning algorithm employing extreme gradient boosting (XGB) is trained to predict  $\varepsilon_c$  by analyzing the optical properties of 160 Ce<sup>3+</sup> substituted inorganic phosphors. The experimentally measured  $\varepsilon_c$  of these compounds was featurized using the materials' relative permittivity ( $\varepsilon_r$ ), average electronegativity, average polarizability, and local geometry. Because the number of reported  $\varepsilon_r$  values is limited, it was necessary to utilize a predicted relative permittivity  $(\epsilon_{r,SVR})$  obtained from a support vector regressor trained on data from ~2,800 density functional theory calculations. The remaining features were compiled from open-source databases and by analyzing the rare-earth coordination environment from each Crystallographic Information File (CIF). The resulting ensemble model could reliably estimate  $\varepsilon_c$  and provides insight into the optical properties of Ce<sup>3+</sup>-activated inorganic phosphors.

Publishing

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

### I. INTRODUCTION

Phosphor-converted white light-emitting diodes (pc-wLEDs) have gained significant attention because they are less toxic than fluorescent lamps and consume less energy than incandescent light bulbs.<sup>1,2</sup> In a typical pc-wLED, a rare-earth substituted inorganic phosphor is excited by blue or near-ultraviolet light from an LED chip. The absorbed photons are then down-converted and re-emitted at longer wavelengths.<sup>3</sup> The resulting combination of the higher energy LED emission and the lower energy phosphor emission covers the entire visible spectrum, thereby appearing as white light. In general, the color quality of this white light is primarily controlled by the phosphor's optical properties, which stem from the rare-earth ion's parity-allowed electric dipole  $5d \leftrightarrow 4f$  transition. The magnitude of the d-f separation is particularly significant because it sets the position of the absorption and emission energies. More importantly, it can be manipulated by varying the interaction between the rare-earth 5d-orbitals and the neighboring anion ligands.

Ce<sup>3+</sup> substituted materials are among the most widely explored rare-earth containing phosphors due to their unique luminescent properties such as broad and highly efficient emission. Typically, the energy gap between the 4f orbitals and the centroid position of the five 5d orbitals for a free Ce<sup>3+</sup> ion in vacuum occurs deep in the ultra-violet region of the electromagnetic spectrum (~6.35 eV; 51,200 cm<sup>-1</sup>). The 5d levels can be stabilized (decreased) in energy relative to the free ion ground state by surrounding the cation with anions, called the nephelauxetic effect. This change can be quantified as the 5d centroid shift ( $\varepsilon_c$ ).<sup>4</sup> Additional energy stabilization can also be achieved by varying the different degrees of covalency and geometry of the cation-anion interaction, known as crystal field splitting ( $\varepsilon_{cfs}$ ).<sup>5–7</sup> The optical properties of a phosphor can be reasonably assessed with the knowledge of the centroid shift and crystal field splitting of the 5d energy levels.

Publishing

This i

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

Studies have revealed the coordination number, polyhedron size, and degree of polyhedron distortion around the Ce<sup>3+</sup> ion control the crystal field splitting of the 4f<sup>n-1</sup>5d-levels, whereas the anion type is generally irrelavent.<sup>8-12</sup> This is in contrast to the centroid shift, which is suggested to be closely associated with the anion ligands. The instantaneous position of Ce<sup>3+</sup> electrons polarize the anion ligands, which results in a self-induced potential that reduces the inter-electron Coulomb repulsion between the Ce<sup>3+</sup> electrons and causes the centroid energy of the 5d levels to lower.<sup>13</sup> The centroid shift is further influenced by the covalency of the bonding within the crystal structure.<sup>14</sup> Indeed, a significant degree of covalency appears to induce a more substantial nephelauxetic effect, and thus, a large value of  $\varepsilon_c$ . For example, the bonds in nitrides and oxynitrides generally exhibit stronger covalency than oxides.<sup>15</sup> As a result, nitride and oxynitride compounds cause a more substantial centroid shift  $\varepsilon_c$ .<sup>16-22</sup>

Considering both the ligand polarization and covalency models,  $\varepsilon_c$  can be estimated following Equation 1 based on the polarizability of *N* coordinating anion ligands at distance  $R_i$  from the rare-earth, in this case, Ce<sup>3+</sup>,<sup>8,9</sup>

$$\varepsilon_c = 1.79 \times 10^{13} \alpha_{sp} \sum_{i=1}^{N} \frac{1}{(R_i - f \Delta R)^6}$$
(1)

where  $\alpha_{sp}$  is the spectroscopic polarizability of the anion,  $\Delta R$  is the ionic radius difference between Ce<sup>3+</sup> and the cation for which it substitutes, and *f* is the correction parameter that accounts for the relaxation of the neighboring anions induced by the rare-earth substitution. The  $\alpha_{sp}$  is a phenomenological parameter that is not only related to the average polarizability of the *N* nearest anionic neighbors around Ce<sup>3+</sup>, but also represents the covalency within the structure. Unfortunately, quantitatively assessing  $\alpha_{sp}$  remains is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Journal of Applied Physics

Publishing

This i

challenging. Thus, it is not always straightforward to predict centroid shift values using this equation.

Relative permittivity ( $\varepsilon_r$ ), also called the dielectric constant, provides information on polarizability and covalency that can be explicitly measured or calculated,<sup>23,24</sup> making it an obtainable proxy for  $\alpha_{sp}$ .<sup>6</sup> In general,  $\varepsilon_r$  varies significantly depending on the material, temperature, frequency of applied field, and other parameters. Nevertheless, it provides information on the type of bonding in a compound. Materials with a small  $\varepsilon_r$  tend to have more ionic-like interactions, such as fluorides, whereas materials with a large  $\varepsilon_r$  tend to be more covalent-like interactions, such as nitrides and sulfides. The value of the relative permittivity can be experimentally obtained from the comparison of the capacity of an airfilled capacitor with that of the same capacitor containing the dielectric substance under specified temperature and frequency. However, these measurements are relatively uncommon, and thus the number of experimentally measured values is small. Fortunately,  $\varepsilon_r$  is an intrinsic material property that can also be estimated using density functional theory (DFT), allowing for the potential of high-throughput implementation.<sup>25</sup> There are still considerable challenges in computationally determining  $\varepsilon_{r}$ ; most notably, calculating  $\varepsilon_{r}$  using DFT is computationally expensive. DFT can also not easily account for atomic disorder, like site sharing, which is common in inorganic phosphor hosts, and it is currently restricted to smaller unit cells containing typically fewer than one hundred atoms. These limitations have permitted  $\varepsilon_r$  to be calculated for only a few thousand ordered crystalline compounds, or <10% of the reported inorganic solids. Luckily, the implementation of machine learning provides an avenue to vastly expand the number of  $\varepsilon_r$  values available with minimal computational cost.

Journal of Applied Physics

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

Machine learning has become a valuable tool for the development of optical materials like inorganic scintillators, inorganic phosphors, and inorganic photovoltaics, among other property-specific problems.<sup>26–32</sup> Embedding machine learning into the physical and chemical sciences not only accelerates the discovery of new materials but also provides insightful knowledge on composition-structure-property relationships. For example, a model based on chemical compositions and electronic structure calculations was recently developed to predict the Coulomb repulsion energy and chemical shift for lanthanide ions in a compound. In combination with the key experimentally measured parameters and the physical-based empirical models, the machine learning model can predict the 4f ground state and 5d<sup>1</sup> excited state electronic configurations of the lanthanide dopant relative to the band edges of the host material, which enables the fast screening for potential candidates in a high-throughput manner.<sup>31</sup>

In this work, we employ a combination of compositional and structural descriptors to build a machine learning regression model that first predicts relative permittivity. This model is trained using ~2,800  $\varepsilon_{r,DFT}$  values extracted from the Materials Project, and allows us to expand our database of  $\varepsilon_{r,SVR}$  values to ~280,000 compositions that are compiled in Pearson's Crystal Database (PCD). We then use a tree-based ensemble learning method implementing extreme gradient boosting (XGB) to predict the centroid shift ( $\varepsilon_{c,XGB}$ ) for Ce<sup>3+</sup>activated inorganic phosphors using a descriptor set including the machine-learning derived relative permittivity as well as average cation electronegativity, average anion polarizability, structure condensation, and Ce<sup>3+</sup> coordination environment. Our successful development of a machine learning method to approximate the centroid shift is a key step to interpret and predict the luminescence properties of inorganic phosphors. Moreover, the centroid shift data combined with luminescence spectroscopy can provide insights into the chemical shift of 4f-electron binding energy.<sup>33</sup>

### II. EXPERIMENTAL

### A. Relative permittivity data extraction and model construction

The machine learning model to predict  $\varepsilon_c$  first requires  $\varepsilon_{r,SVR}$  as a descriptor. The  $\varepsilon_{r,DFT}$  of 4,867 compounds were therefore extracted from the Materials Project database.<sup>25</sup> The Materials Project calculates  $\varepsilon_{r,DFT}$  using a high-throughput framework based on Density Functional Perturbation Theory (DFPT).<sup>34–36</sup> In some cases, negative values of  $\varepsilon_{r,DFT}$  are present in the database; these were removed from the final training dataset. The number of compositions was further reduced to 2,991 after cross-referencing with PCD<sup>37</sup> to ensure that all the phases used for machine learning are also experimentally reported, i.e. hypothetical crystal structures were removed. The data distribution plot is provided in Figure S1. About 5% of the data possess an  $\varepsilon_{r,DFT} > 20$  and these tend to be transition metal-containing compositions unsuitable as phosphor hosts. Thus, these compounds were also removed by fixing the learning window to  $0 < \varepsilon_{r,DFT} < 20$ . Compounds with elements from group 18 (noble gases) and Z > 83 (except for U and Th) were also excluded. These criteria reduced the final training set to 2,832 compounds.

The descriptors used in this study included 17 distinct compositional variables describing elemental properties such as position on the periodic table, electronic structure, and physical properties as well as their associated math expressions (difference, average, largest value, smallest value, and standard deviation). Additionally, 13 structural descriptors related to variables, including crystal system, space group, and unit cell volume, among others, were incorporated in the machine learning algorithm. The full list of descriptors and

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

This i

or Physics

Journal of Applied Ph

Publishing

This i

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

mathematical expressions is available in the Supporting Information (Table S1). In total, 98 descriptors were used to build the model.

The 2,832 × 98 data matrix was split into a random train and test subsets with a 9:1 ratio. The training set was then standardized to have a mean of 0 and a variance of 1, and the test set was transformed using the same scaler as the training set. The machine learning model was constructed using a Support Vector Regression (SVR) algorithm with the radial basis function (RBF) kernel.<sup>38</sup> A grid search, which exhaustively evaluates all parameter combinations, was performed on the training set with a 10-fold cross-validation method to choose the best hyper-parameter settings. The searching space was defined as cost (C) values ranging in log space [0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2], and epsilon values in real space [0.001, 0.01, 0.1, 1], where cost is the regularization term and epsilon specifies the epsilon-tube within which no penalty is associated in the training loss function. The scikit-learn python implementations of these learning algorithms were used.<sup>39</sup>

### B. Centroid shift data extraction and model construction

The development of the machine learning model to predict the centroid shift of Ce<sup>3+</sup>activated phosphors involved extracting 219 experimentally reported host compositions and the associated  $\varepsilon_c$  values from the literature.<sup>5,7–11,40</sup> Before training, data were examined according to two primary criteria. First, phosphors hosts that do not have a reliable crystal structure reported were omitted. The second criterion is that the Ce<sup>3+</sup> must have a single, chemically obvious substitution site or a site that was specifically studied and reported in the paper if the compound contains multiple substitution sites. The final number of training labels was reduced to 160 phosphors. Eight features, including relative permittivity, which was predicted using the first  $\varepsilon_{r,SVR}$  model, weighted average cation electronegativity, weighted average anion polarizability, and local environment related descriptors, were

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

This i

employed to characterize these phosphors. A train-test splitting ratio of 9:1 was also used to randomly split the data into two subsets.

The tree-based extreme gradient boosting (XGB) algorithm was selected to train the centroid shift model owing to its effectiveness in obtaining satisfactory results with small datasets while controlling overfitting.<sup>41</sup> The model was gradient boosted for 50 rounds with a learning rate of 0.15 and a maximal tree depth of 3. These actions were carried out using the scikit-learn API.<sup>39</sup>

### C. Data availability

The training and test datasets, as well as the codes used to generate the models in this work can be downloaded from https://github.com/BrgochGroup/CentoidShiftPredictor.

### III. RESULTS AND DISCUSSION

### A. Modeling relative permittivity

The objective of SVR is to find a multi-dimentional function, f(x), that deviates from the training label by a value no greater than epsilon ( $\varepsilon$ ) for each training sample and is simultaneously as flat as possible. The x in f(x) is a set of feature vectors that represent the compounds and the solution of f(x) is the predicted centroid shift value. A regularization parameter, C, also called cost, is introduced to control the compromise between model accuracy and flatness. As C increases, the strength of the regularization increases and then the tolerance for data outside of  $\varepsilon$  increases. As C approaches 0, the tolerance approaches 0 and the equation collapses into the simplified one. The two parameters, C and  $\varepsilon$ , must first be optimized to maximize the model performance. Conducting a grid search for the free parameters, provided in Figure S2, shows that values of  $C = 10^{1.75}$  ( $\approx$ 56.23) and  $\varepsilon = 0.1$  give the best model performance based on the model statistics, i.e., coefficient of determination  $(r^2)$  and mean absolute error (MAE).



Figure 1. Data distribution of 2,832 training samples that were calculated with DFT. The bars represent the total counts in each bin, and the curve represents the cumulative percentage.

The  $\varepsilon_r$  was trained using these parameters and 2,832 training labels collected from the Materials Project. Figure 1 shows the training data distribution. Over half of the compositions have a DFT-calculated relative permittivity ( $\varepsilon_{r,DFT}$ ) below 5, and  $\approx$ 95% have an  $\varepsilon_{r,DFT} \leq 12.5$ . The SVR model was evaluated with a test set, which was obtained from randomly selecting 10% data before training and completed unseen by the model. As shown in Figure 2, the histograms at the top and right show that both the DFT-calculated relative permittivity ( $\varepsilon_{r,DFT}$ ) and the machine learning predicted values ( $\varepsilon_{r,SVR}$ ) in the test set represent the distribution of the entire training labels well. Acceptable agreement is obtained between  $\varepsilon_{r,DFT}$  and  $\varepsilon_{r,SVR}$  with the coefficient of determination ( $r^2$ ), and mean absolute error (MAE) being 0.93 and 0.65, respectively. 246 out of 284 (87%) compounds in the test set were predicted with an error of less than 25%. There is a slight but noticeable underestimation for 12.5 <  $\varepsilon_r \leq$  20, which is most likely due to the limited number of training samples in this region. Nevertheless, this model is useful to estimate the  $\varepsilon_{r,SVR}$  for over 270,000 compounds compiled in PCD.

ACCEPTED MANUSCRIPT

the online version of record will be different from this version once it has been copyedited and typeset.

PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

is the author's peer reviewed, accepted manuscript. However,

Journal of Applied Physic

Publishing



Figure 2. Predicted relative permittivity ( $\epsilon_{r,SVR}$ ) versus DFT-calculated relative permittivity ( $\epsilon_{r,DFT}$ ) is shown for a 10% holdout test set. The ideal line is shown as the dashed gray line and the fit line is shown as the solid yellow line. The curves at the top and right show the histograms of the data.

### B. Connecting relative permittivity to the centroid shift prediction

To build the centroid shift model, 219 experimentally measured centroid shift ( $\epsilon_c$ ) data were collected from literature for Ce<sup>3+</sup>-activated inorganic phosphors. Data reduction was then performed. For example, although the  $\epsilon_c$  data are available for BaCaBO<sub>3</sub>F and Sr<sub>2</sub>SiO<sub>4</sub>, they were excluded because there is no clear information on the Ce<sup>3+</sup> substitution site or the associated optical properties stemming from substitution on each site. Indeed, the obtained value of  $\epsilon_c$  for these systems could stem from Ce<sup>3+</sup> occupying several crystallographic sites. This ambiguity can cause issues in machine learning; thus, these data were removed. Data sanitization reduced the final training set to 160 data, including 53 (oxy)halides, 85 oxides, 5 sulfides, 2 selenides, and 15 nitrides. All of these phosphors contain a single substitution site, or if multiple crystallographic sites are present, the rare-earth site occupancy was specifically investigated and reported in the literature. The data are provided in Table S2.

10

the online version of record will be different from this version once it has been copyedited and typeset

PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

is the author's peer reviewed, accepted manuscript. However,

AIP Publishing Applied Physic

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

Analyzing these data by plotting the  $\varepsilon_c$  against the type of anions present in the composition (Figure 3a) demonstrates that the centroid shift tends to follow some trends. Fluorides have the smallest  $\varepsilon_c$  as dictated by the ionic bonding of these phases, whereas sulfides, selenides, and nitrides possess a relatively larger  $\varepsilon_c$  values resulting from their more covalent bonding. Although a crude estimation of  $\varepsilon_c$  can be made based on the anions present in the host compounds,  $\varepsilon_c$  may still vary greatly for compounds that are composed of the same type of anion. This variation is particularly noticeable for oxides and nitrides. One possible explanation is that the susceptibility of O<sup>2-</sup> and N<sup>3-</sup> to the electronegativity of the surrounding cations is larger than anion such as F<sup>-</sup>. In other words, the anion polarizability and covalency are affected more strongly by the cations in oxides and nitrides, leading to greater variability in the properties. This observation is a bit unfortunate because these compound types are among the most common classes of inorganic phosphor hosts, yet, they have the biggest range of centroid shifts.

The same general trend was also observed in the plot of  $\varepsilon_{r,SVR}$  against the type of anions. Because the experimentally measured or DFT calculated relative permittivity for these 160 compositions are unknown, the  $\varepsilon_{r,SVR}$  model was used for this comparison. As shown in Figure 3b, the  $\varepsilon_{r,SVR}$  in general shows a trend as  $I^- > Br^- > CI^- > F^-$  in halides, and sulfides, selenides, and nitrides possess a larger  $\varepsilon_{r,SVR}$  than oxides. These trends agree well with those in  $\varepsilon_c$ , as expected. There is a slightly larger spread in this plot, however, meaning that using  $\varepsilon_{r,SVR}$  solely to predict  $\varepsilon_c$  is subject to error. For example, although LaPO4:Ce<sup>3+</sup> and GdAlO<sub>3</sub>:Ce<sup>3+</sup> have a very close  $\varepsilon_{r,SVR}$  (4.23 and 4.25, respectively), they show a large discrepancy in  $\varepsilon_c$  ( $\varepsilon_c = 1.07$  eV for LaPO4:Ce<sup>3+</sup> and  $\varepsilon_c = 1.70$  eV for GdAlO<sub>3</sub>:Ce<sup>3+</sup>).<sup>7,42,43</sup> The failure stems from the fact that a simple descriptor like  $\varepsilon_{r,SVR}$  is not sufficient to account for all of the impacting factors of the centroid shift, such as the size effect. For instance,

permittivity can be expressed as a sum of contributions from the different atoms in a compound, and a large cation, such as Cs<sup>+</sup>, takes a large volume of the lattice, but it may not contribute much to the permittivity. However, its large size most likely leads to a smaller  $\varepsilon_c$ , according to Equation 1.



Figure 3. (a) The centroid shift ( $\varepsilon_c$ ) and (b) the machine learning obtained relative permittivity ( $\varepsilon_{r,SVR}$ ) for Ce<sup>3+</sup> inorganic compounds displayed against the type of anions. The darker regions represent a higher density of data points.

### C. Machine learning the centroid shift

The ability to predict the centroid shift is clearly multi-dimensional. An ensemble learning method was therefore constructed to predict  $\varepsilon_c$  accurately with features chosen based on numerical equations for centroid shift and the crystal structures. The features and the corresponding notations used in this work are provided in Table 1, and includes  $\varepsilon_{r,SVR}$ , average cation electronegativity defined in Equation 2,

$$\chi_{av} = \frac{\sum n_i z_i \chi_i}{\sum n_i z_i}$$
(2)

where  $n_i$  is the stoichiometric index of cation i with charge  $z_i$  and  $\chi_i$  is the Pauling electronegativity<sup>44</sup>. The average anion polarizability is calculated in Equation 3,

$$\alpha_{av} = \frac{\sum m_e \alpha_e}{\sum m_e} \tag{3}$$

12

manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434 Journal blied Publishing

is the author's peer reviewed, accepted manuscript.

where  $m_e$  is the stoichiometric index of anion e, and  $\alpha_e$  is the polarizability. In addition,  $\Delta R$  is the difference between the ionic radius compiled by Shannon of Ce<sup>3+</sup> and the cation (R<sub>M</sub>) for which Ce<sup>3+</sup> substitutes.<sup>45</sup> The final descriptor is condensation (cond.), which is the ratio between the number of anions and the number of cations contained in the chemical formula.

Table 1. The feature set used to predict the centroid shift modeled using XGB.

Variable		Notation
Relative permittivity		ε <sub>r,SVR</sub>
Average cation e	lectronegativity	χav
Average anion polarizability		α <sub>av</sub>
Ionic radius		R <sub>M</sub>
Difference in rad	ius	$\Delta R$
Average bond length		R <sub>av</sub>
Coordination nu	mber	Ν
Condensation		cond.

The model was subsequently trained using a tree-based XGB algorithm with 10% of the data held out for model evaluation. To account for randomness in the test data selection, ten different test sets were examined, which yielded an average  $r^2$  of 0.90. Figure 4a shows the results from one of these test sets. The detailed statistics are provided in Table S3. The statistics of the test set prediction show compelling agreement between the  $\varepsilon_{c,exp}$ , and the  $\varepsilon_{c,XGB}$  with a root-mean-squared error (RMSE) of 0.18 eV and a mean absolute error (MAE) of 0.13 eV. These results indicate that the feature set does a reasonable job determining a Ce<sup>3+</sup> phosphor's centroid shift.

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

This i



Figure 4. (a) Machine learning predicted centroid shift ( $\varepsilon_{c,XGB}$ ) against the experimentally measured centroid shift ( $\varepsilon_{c,exp}$ ). The test set plotted here represents 10% of the training set. The ideal line is shown as the dashed gray line and the fit line is shown as the solid red line. (b) Feature importance of the XGB model in terms of gain and coverage.

Analyzing the feature importance in this model can provide insight into the crystalchemical properties that control the centroid shift. Different feature importance matrices are accessible from the XGB model. The principle of developing a tree is that a new split on a node will be added only if the split results in a more accurate prediction. The gain can quantify the improvement in accuracy due to the split of a specific feature at a node. The gain in feature importance is the total gain across all splits in which the feature is used, implying the relative contribution of the corresponding feature to the model. As shown in the top panel of Figure 4b, the average anion polarizability ( $\alpha_{av}$ ) has a significantly higher absolute gain than all other features demonstrating it is the most important feature for predicting  $\varepsilon_c$ . Moreover, relative permittivity and average cation electronegativity ( $\chi_{av}$ ) also show a non-negligible influence on the model performance, while the remaining five features have a minimal total gain. Although the remaining five features have low gain, removing them from the model and re-training leads to considerably worse performance. All of the descriptors are essential to obtain  $\varepsilon_{c,XGB}$ .

the online version of record will be different from this version once it has been copyedited and typeset.

PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

However,

is the author's peer reviewed, accepted manuscript.

This

Journal of Applied Physic

Publishing

Journal of Applied Physics Publishing

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434 This

Feature importance can also be evaluated based on the coverage matrix, which describes the relative quantity of samples related to a feature. Here, the trend is slightly different. As shown in the bottom panel of Figure 4b, the coverage matrix suggests that  $\epsilon_{r,SVR}$  is the most crucial feature in the coverage matrix. More than 5,000 observations were associated with the split of  $\epsilon_{r,SVR}$  in the boosting process. The  $\alpha_{av}$  and  $\chi_{av}$  relate to ~2,500 observations, while  $R_M$  is the least important although it still influences 989 observations. Overall,  $\epsilon_{r,SVR}$ ,  $\chi_{av}$ , and  $\alpha_{av}$  are the three most essential features related to the centroid shift regardless of the feature importance matrix. Not surprisingly, the average anion polarizability is more correlated to the centroid shift than the average cation electronegativity. The local geometry of the luminescent center and condensation contribute less to the model yet, they are still necessary to predict  $\epsilon_c$ .

It is vital to quantitatively estimate the position of the 5d-excited states for rareearth ions in host crystal structures for designing inorganic phosphors. A phosphor with a large centroid shift tends to have a longer wavelength emission and using this prediction of centroid shift makes it possible to roughly estimate the emission color. The ability to predict  $\varepsilon_c$  even with moderate quantitative accuracy will allow researchers to accelerate phosphor discovery by performing a top-level screening of new phosphor hosts with desired optical properties. Moreover, the  $\varepsilon_c$  model can be combined with the knowledge of crystal field splitting and the phosphor host's bandgap to predict the location of 5d levels with respect to the bottom of the conduction band.<sup>40</sup> This combination would provide insight into the thermal quenching behavior of the phosphor that has thus far remained largely empirical.

### **IV. CONCLUSIONS**

Combining high-throughput DFT calculations and machine learning techniques provides a unique framework to predict the relative permittivity of over 270,000

Journal of Applied Physics Publishing

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434 This

compounds compiled in inorganic structure databases. The predicted relative permittivity was then combined with chemical properties and local geometry of 160  $Ce^{3+}$ -activated phosphors to predict the centroid shift. Ensemble learning methods were successfully employed to predict the values of the centroid shift, which is beneficial to interpret and predict the luminescence properties and thermal quenching behaviors for Ce<sup>3+</sup>-doped phosphors. The results combined with the information of crystal field splitting can also be extended into the estimation of properties of other rare-earth doped materials, for example, the Eu<sup>2+</sup> inter 4f-electron Coulomb repulsion energy in compound. In addition, the predicted  $\varepsilon_{c}$  model serves as a starting point to generate the binding energy of electrons at the top of the valence band. This information is not only of interest for luminescence but also for many other disciplines of science such as photo-catalytic splitting of water, battery potentials, valence band offsets in semi-conductor hetero-junction or in core shell particles among applications. Further tuning and improvement of these models for  $\varepsilon_c$  has the potential to make a broad impact on assisting the development of numerous functional inorganic materials.

### SUPPLEMENTARY MATERIAL

See supplementary material for the following: Table S1. Feature set of the relative permittivity model; Table S2. Centroid shift training lables; Table S3. Centroid shift model statistics of ten randomly generated test sets; Figure S1. Relative permittivity data distribution; Figure S2. Plot of parameter optimization for cost and epsilon.

### ACKNOWLEDGMENTS

The authors thank the National Science Foundation (DMR 18-47701). This work was also supported by the Welch Foundation (E-1981) and the Texas Center for Superconductivity at the University of Houston (TcSUH). The authors declare no competing financial interest.

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

### REFERENCES

<sup>1</sup> J.K. Kim and E.F. Schubert, Opt. Express **16**, 21835 (2008).

<sup>2</sup> S. Pimputkar, S.J. Speck, P.S. DenBaars, and S. Nakamura, Nat. Photonics **3**, 180 (2009).

<sup>3</sup> N.C. George, K.A. Denault, and R. Seshadri, Annu. Rev. Mater. Res. 43, 481 (2013).

<sup>4</sup> C. Jørgensen, *Modern Aspects of Ligand Field Theory* (North-Holland Publishing Company,

Amsterdam, The Netherlands, 1971).

<sup>5</sup> P. Dorenbos, J. Lumin. **135**, 93 (2013).

<sup>6</sup> M. Mikami and N. Kijima, Opt. Mater. (Amst). **33**, 145 (2010).

<sup>7</sup> P. Dorenbos, Phys. Rev. B **65**, 235110 (2002).

<sup>8</sup> P. Dorenbos, Phys. Rev. B **62**, 15640 (2000).

<sup>9</sup> P. Dorenbos, Phys. Rev. B **62**, 15650 (2000).

<sup>10</sup> P. Dorenbos, Phys. Rev. B **64**, 125117 (2001).

<sup>11</sup> P. Dorenbos, J. Lumin. **99**, 283 (2002).

<sup>12</sup> P. Dorenbos, J. Alloys Compd. **341**, 156 (2002).

<sup>13</sup> C.A. Morrison, J. Chem. Phys. **72**, 1001 (1980).

<sup>14</sup> B.F. Aull and H.P. Jenssen, Phys. Rev. B **34**, 6640 (1986).

<sup>15</sup> F.J. DiSalvo and S.J. Clarke, Curr. Opin. Solid State Mater. Sci. **1**, 241 (1996).

<sup>16</sup> T. Wang, Q. Xiang, Z. Xia, J. Chen, and Q. Liu, Inorg. Chem **55**, 2929 (2016).

<sup>17</sup> T. Suehiro, N. Hirosaki, and R.J. Xie, ACS Appl. Mater. Interfaces **3**, 811 (2011).

<sup>18</sup> J.W.H. Van Krevel, J.W.T. Van Rutten, H. Mandal, H.T. Hintzen, and R. Metselaar, J. Solid State Chem. **165**, 19 (2002).

<sup>19</sup> Y.Q. Li, A.C.A. Delsing, G. De With, and H.T. Hintzen, Chem. Mater. **17**, 3242 (2005).

<sup>20</sup> L. Zhang, J. Zhang, X. Zhang, Z. Hao, H. Zhao, and Y. Luo, ACS Appl. Mater. Interfaces **5**, 12839 (2013).

This i

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

This i

<sup>21</sup> S. Schmiechen, H. Schneider, P. Wagatha, C. Hecht, P.J. Schmidt, and W. Schnick, Chem. Mater. **26**, 2712 (2014).

<sup>22</sup> W.Y. Huang, F. Yoshimura, K. Ueda, Y. Shimomura, H.S. Sheu, T.S. Chan, C.Y. Chiang, W. Zhou, and R.S. Liu, Chem. Mater. **26**, 2075 (2014).

<sup>23</sup> S. Ishihara, M. Tachiki, and T. Egami, Phys. Rev. B **49**, 16123 (1994).

<sup>24</sup> S.H. Wemple and M. DiDomenico, Phys. Rev. B **3**, 1338 (1971).

<sup>25</sup> A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, APL Mater. **1**, 011002 (2013).

<sup>26</sup> G. Pilania, X.Y. Liu, and Z. Wang, J. Mater. Sci. **54**, 8361 (2019).

<sup>27</sup> Y. Zhuo, S. Hariyani, E. Armijo, Z. Abolade Lawson, and J. Brgoch, (2019).

<sup>28</sup> S. Ganguly, C.S. Kong, S.R. Broderick, and K. Rajan, Mater. Manuf. Process. **28**, 726 (2013).

<sup>29</sup> B.J.M. Webb-Robertson, K.F. Ferris, and D.M. Jones, IEEE Trans. Nucl. Sci. 55, 1210 (2008).

<sup>30</sup> H. Jin, H. Zhang, J. Li, T. Wang, L. Wan, H. Guo, and Y. Wei, J. Phys. Chem. Lett **11**, 3075 (2020).

<sup>31</sup> G. Pilania, K.J. McClellan, C.R. Stanek, and B.P. Uberuaga, J. Chem. Phys. **148**, 241729 (2018).

<sup>32</sup> Y. Zhuo, A. Mansouri Tehrani, A.O. Oliynyk, A.C. Duke, and J. Brgoch, Nat. Commun. **9**, 4377 (2018).

<sup>33</sup> P. Dorenbos, Phys. Rev. B **85**, 165107 (2012).

<sup>34</sup> I. Petousis, W. Chen, G. Hautier, T. Graf, T.D. Schladt, K.A. Persson, and F.B. Prinz, Phys. Rev. B **93**, 115151 (2016).

<sup>35</sup> I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T.D. Schladt,
 K.A. Persson, and F.B. Prinz, Sci. Data 4, 160134 (2017).

<sup>36</sup> S. Baroni, P. Giannozzi, and A. Testa, Phys. Rev. Lett. **58**, 1861 (1987).

is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset. PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0012434

Publishing

This i

<sup>37</sup> O. Villars and K. Cenzual, *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds* (ASM International, Ohio, USA, 2007).

<sup>38</sup> H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, Adv. Neural Inf. Process. Syst. **1**, 155 (1996).

<sup>39</sup> F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

<sup>40</sup> S. Wang, Z. Song, Y. Kong, Z. Xia, and Q. Liu, J. Lumin. **200**, 35 (2018).

<sup>41</sup> T. Chen and C. Guestrin, in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* (New York, New York, USA, 2016), pp. 785–794.

<sup>42</sup> J.W.M. Verweij, M.T. Cohen-Adad, D. Bouttet, H. Lautesse, B. Moine, and C. Pédrini, Chem.
Phys. Lett. 239, 51 (1995).

<sup>43</sup> M. V Hoffman, J. Electrochem. Soc. **118**, 1508 (1971).

<sup>44</sup> L. Pauling, *The Nature of the Chemical Bond*, 3rd ed. (Cornell University Press, Ithaca, New York, 1960).

<sup>45</sup> R.D. Shannon, Acta Crystallogr. Sect. A **32**, 751 (1976).



# ACCEPTED MANUSCRIPT



AIP

### Journal of Applied Physics

# ACCEPTED MANUSCRIPT





# ACCEPTED MANUSCRIPT





# ACCEPTED MANUSCRIPT

