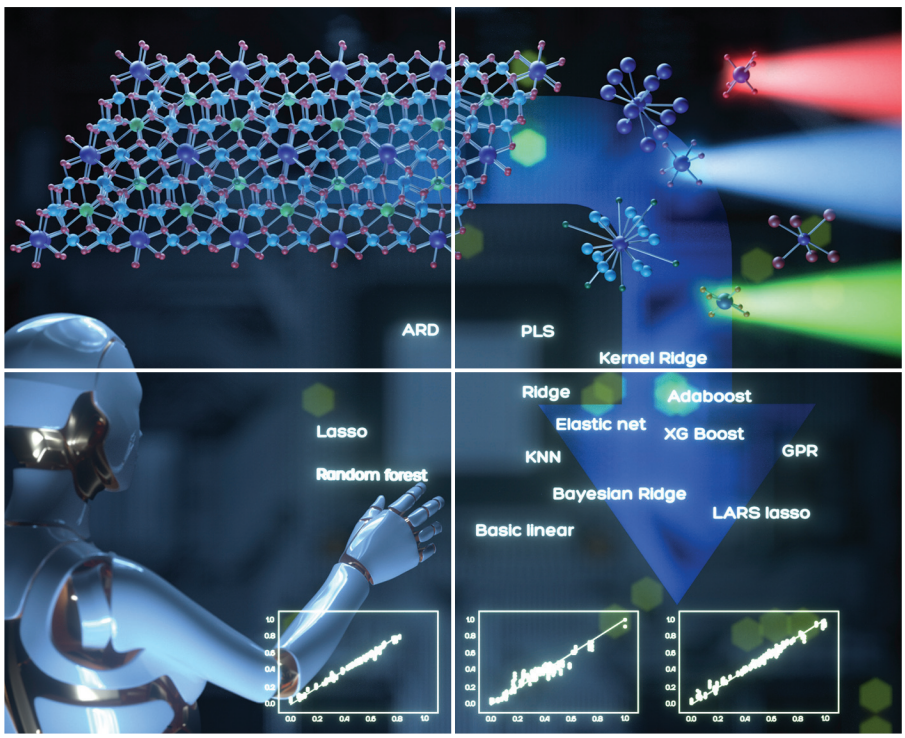
Volume 8 | Number 21 | 7 November 2021



INORGANIC CHEMISTRY

FRONTIERS

CCS 中国化学会 CHINESE CHEMICAL SOCIETY
PEKING UNIVERSITY 1898
ROYAL SOCIETY OF CHEMISTRY
rsc.li/frontiers-inorganic

INORGANIC CHEMISTRY FRONTIERS

RESEARCH ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Inorg. Chem. Front.*, 2021, **8**, 4610

A data-driven approach to predicting band gap, excitation, and emission energies for Eu^{2+}-activated phosphors†

Chaewon Park,‡[a] Jin-Woong Lee,‡[a] Minseuk Kim,‡[a] Byung Do Lee,[a] Satendra Pal Singh,[a] Woon Bae Park*[b] and Kee-Sun Sohn *[a]

The prediction of excitation band edge wavelength (EBEW) and peak emission wavelength (PEW) for Eu^{2+}-activated phosphors is intricate in practice, although a theoretical interpretation has been well established. A data-driven approach could be of great help for EBEW and PEW prediction. We collected 91 Eu^{2+}-activated phosphors, the host structures of which exhibit a single activator site and the EBEW and PEW of which are available at the critical activator concentration. We extracted 29 descriptors (input features) that implicate the elemental and structural traits of phosphor hosts, and set up an integrated machine-learning (ML) platform consisting of 18 ML algorithms that allowed prediction of the EBEW and PEW as well as the DFT-calculated band gap (E_g). The acquired dataset involving 91 phosphors was insufficient for the 29-input-feature problem and the real-world data collected from the literature have a so-called dirty nature due to inaccurate, unstandardized experiments. Despite an unavoidable paucity of data and the dirty-data problems of real-world data-based ML implementation, we obtained acceptable holdout dataset test results for PEW predications such as $R^2 > 0.6$, MSE < 0.02, and test_R^2/training_R^2 > 0.77 for four ML algorithms. The EBEW and E_g predictions returned slightly better test results than these PEW examples.

Received 18th June 2021,
Accepted 19th August 2021

DOI: 10.1039/d1qi00766a

rsc.li/frontiers-inorganic

Introduction

Phosphors play a crucial role in light-emitting diode (LED) applications, and many Ce^{3+}, Eu^{2+}, and Mn^{4+}-activated novel LED phosphors have been consistently discovered.[1–10] Data-driven approaches have recently been mainstays in the field of phosphor research to facilitate phosphor discovery.[11–20] The first data-driven approach[11] to inorganic phosphors was initiated in 2015 by the present authors, wherein confirmatory factor analysis was employed to predict the peak emission wavelength for Eu^{2+}-activated phosphors. Thereafter, other promising data-driven approaches have been reported.[12–16] A monumental ML-based phosphor discovery was reported by Brgoch *et al.*,[16] wherein a ML technique (support vector machine) in conjunction with a materials project database[17] and DFT calculations was used for the Debye temperature prediction of 2071 phosphor hosts, which eventually led to the discovery of a brilliant novel phosphor, $NaBaB_9O_{15}$:Eu^{2+}, which is the first example of the use of DFT calculations and ML algorithms for data-driven phosphor discovery. Brgoch *et al.*,[18] thereafter, improved the ML portion by introducing an XG boost algorithm. In the meantime, a series of well-organized data-driven phosphor predictions have also been well established by Ong *et al.*[12–15] Brik *et al.*[19] very recently reported the use of a semi-data-driven approach using a basic linear regression along with an appropriate knowledge-based feature selection, and revealed a relationship between the structural properties of the hosts and the optical properties of the Eu^{2+} dopant. In addition to phosphor research, there have been many more advances in data-driven materials discovery approaches in other inorganic science areas.[21–24]

The ML approach to phosphor research has practical challenges such as real-world data paucity, scattered data distribution due to inconsistent experimental settings (the dirty nature of real-world data), feature (descriptor) selection complexity, and a difficulty in relevant algorithm selection. There is a serious lack of real-world data for inorganic phosphors, and standardized, labeled data acquisition is far from complete. In such situations, domain knowledge allows the ML

[a]*Nanotechnology & Advanced Materials Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul, 143-747, South Korea. E-mail: kssohn@sejong.ac.kr*
[b]*Department of Printed Electronics, Sunchon National University, 291-19 Jungang-ro, Sunchon, Chonnam, 540-742, South Korea. E-mail: wbpark@scnu.ac.kr*

†Electronic supplementary information (ESI) available: Supplementary Fig. S1–S4, and Tables S1–S7. See DOI: 10.1039/d1qi00766a

‡These authors contributed equally.

approach to achieve a more reasonable level of data acquisition. In this regard, specific knowledge of phosphor physics and chemistry was employed for the data acquisition, and this differs from the blind gathering of irrelevant data. For instance, the phosphor data used for the present ML approach was confined to Eu^{2+}-activated phosphors with a single activator site and further limited to the excitation band edge wavelength (EBEW) and peak emission wavelength (PEW) data measured at a critical Eu^{2+} concentration that must be obtained from concentration-quenching data (= phosphor performance data with respect to the activator concentration). Notwithstanding the lack of real data for the band gap of the host (E_g), we collected the appropriate E_g values by employing density functional theory (DFT) calculations and the materials project database.[17] As a result, only 91 phosphors have been collected. As far as we could ascertain, these entries are the only cases, as the list was narrowed down from more than 10 000 papers dealing with Eu^{2+}-activated phosphors.

Brgoch *et al.*[16] sorted out the data paucity problem by employing a massive amount of DFT-generated data for inorganic compounds (including non-phosphor materials) residing in a well-established database as training data and further expanding the fully trained ML model to a small phosphor dataset. This could be considered a brilliant case of transfer learning, which currently is a booming trend in applications of ML.[25] In addition, a successful transfer-learning example[26] was reported based on DFT-based formation energy data collected from an open quantum materials database (OQMD).[27] However, we aimed to develop a small ML model by focusing only on a very specific phosphor group in a narrow range, *viz.*, phosphors with a single-activator site and those with EBEW and PEW data gleaned from the concentration-quenching data. No previous dataset has included EBEW and PEW, although a massive amount of collected data exists for DFT-calculated Debye temperature, formation energy, band gap, elastic constant, dielectric constant, and many other thermodynamic variables,[17,27,28] all of which seem impossible to be directly connected to the EBEW and PEW of phosphors. In addition to the problems of a paucity of real-world data and the intractability of transfer learning, the data acquired thus far has a so-called 'dirty nature' due to different experimental setups and to human intervention during production, which leads to data that is non-identical and independently distributed (non-IID).[29]

Adequate selection of the descriptors that refer to the elemental, structural, physical, and chemical nature of phosphor hosts is a key issue for successful ML modeling for E_g, EBEW, and PEW prediction. The local structure of inorganic compounds has been expressed using various descriptors.[12–16,30–33] Ong *et al.* has used systematic math to succinctly summarize logic-based descriptors,[12–15] and Brgoch *et al.*[16] have proposed 150 descriptors for their ML modeling. Takemura *et al.*[34] have very recently reported a brilliant metric for the dissimilarity measure of local structure, which is based on the Wasserstein distance. In addition, Xie and Grossman[32] have recently used graph theory to devise a brilliant descriptor for the local structures of inorganic materials. The number of descriptors should be closely associated with the size of the training dataset. Since our training dataset includes only 91 phosphors, the number of descriptors had to be relative to this small dataset size. Accordingly, we had to reduce the number of descriptors and finally pinpointed the 29 descriptors representing elemental, structural, physical, and chemical information. Principal-component analysis (PCA), Pearson-, and Spearman-correlation analyses were performed for the dataset, which validated the suggested descriptor selection.[11] Nonetheless, 29 descriptors are still too many when considering the data paucity (91 samples only).

Selection of the ML algorithm should also be relative to the size of the problem as well as to the size of the dataset. The number of descriptors (input features) and target variables (output features) can be used to estimate the size of the problem. For our relatively small dataset that includes 91 phosphors, we had 29 input features and an output feature. Under these circumstances, regularization techniques should definitely be introduced. First, we employed regularized linear regression algorithms such as ridge,[35] Lasso,[36] elastic net,[37] kernel ridge,[38] least-angle regression (LARS) Lasso,[39] Bayesian ridge,[40] and automatic relevance determination (ARD)[41] regressions. We also adopted ensemble algorithms such as random forest,[42] Ada boost,[43] gradient boost,[44] and XG boost.[45] Furthermore, k-nearest neighbor (KNN),[46] support vector machine (SVM),[47] Gaussian process regression (GPR),[48] and partial least square (PLS)[49] were also employed. In addition, a typical artificial neural network (ANN)[50] along with an ordinary linear regression were adopted as baseline methods. Rather than a single or, at best, a few regression algorithms in a conventional ML approach, almost all possible regression algorithms were introduced in the present investigation. There have been reported similar approaches using several ML algorithms simultaneously for a single problem in some other materials research society, *e.g.*, metallic alloys.[51–54] We refer to this sort of approach as an 'integrated ML platform', which can be recommended for phosphor researchers who suffer from real-world data paucity problems.

Experimental (computational details)

ML model selection

We introduced 18 ML algorithms in three categories. The first category includes regularization-based linear regressions such as ridge regression, least absolute shrinkage, and selection operator (Lasso) regression, least-angle regression (LARS), elastic net regression (ENR), kernel ridge regression (KRR), Bayesian ridge regression (BRR), and Bayesian automatic relevance determination (ARD). L2 (ridge) and L1 (Lasso) regularizations were preferably adopted, and other supplementary methodologies such as LARS, KRR, BRR, and ARD were also incorporated. The LARS algorithm exploits the special structure of the Lasso problem, and provides an efficient way to simultaneously compute the solutions for all values of weights. Several kernels such as linear, polynomial, radial

basis function (RBF), sigmoid, and matern kernels were incorporated in the KRR (and also in SVR) to account for the nonlinearity. The Bayesian approach was introduced in the BRR and ARD (and also in GPR) so that the final prediction was distributive rather than deterministic in these cases. The prediction is made with a certain mean and variance, and the same was true for the fitted parameters (= weights).

The second category consists of ensemble algorithms, which create a final model based on a collection of individual models. The predictability of these individual models was weak and could have likely led to over-fitting, but combining many such weak models in an ensemble led to a much improved prediction. We employed several tree ensemble methods such as random forest (RF), adaptive (Ada) boost, gradient boost, and extreme gradient (XG) boost regressions. The results from two more boost algorithms are not presented here since no conspicuous improvement was detected. There are two representative model implementations in an ensemble. The bagging for use in RF treats each model independently, and the boosting from Ada, gradient, and XG boost algorithms sequentially treats each model by putting more weight on the data and on features involving the wrong predictions and high rates of error. The base estimator (each weak model) from a boosted ensemble is a decision tree with a certain depth according to the chosen algorithm (*e.g.*, Ada boost only uses stumps and not trees).

We also incorporated some other well-known regression algorithms such as support vector machine regression (SVR), k-nearest neighbors (KNN), partial least square (PLS), and Gaussian process regression (GPR). More interestingly, the ANN was also employed for the PEW prediction despite the extreme paucity of data, which led to an extreme amount of over-fitting. SVR had been widely used and tuned for versatility to sort out many problems before deep learning was used. The use of kernels is essential in SVR and an RBF kernel was selected from the hyper-parameter optimization process in the present SVR implementation. KNN is the simplest ML algorithm on earth, but the *ad hoc* determination of appropriate k values is essential. PLS is a traditional regression method that is even applicable for a problem with fewer data points than the number of input features, which is a situation very similar to ours. Our problem was ameliorated, however, by a number of data points (91) that was higher than the number of input features (29). GPR, so-called kriging, has recently gained popularity and is often used as a surrogate function for Bayesian optimization.[55] In particular, it is worthwhile to note KNN and GPR since these are parameter-free ML algorithms that differ from most ML algorithms that are concerned with a search for optimal parameters constituting a mathematical model such as a linear model and ANN. Fig. 1 succinctly describes the entire ML platform. All the above-described regression algorithms are available in the Scikit-learn module[56] with well-established default hyper-parameters, none of which were incorporated here, however. We performed an additional hyper-parameter optimization process, which will be discussed in the following subsection.

Training, validation, and test dataset splitting

As mentioned repeatedly, the substantial problem we faced had to do with the paucity of data, which is why we dealt with all the regularization-involved ML algorithms. Under such a situation of insufficient data, special care should be taken when splitting the data into training, validation, and test datasets. Because of the small dataset size, only a simple split into training and test datasets was not viable. We adopted three training schemes. First, we adopted a 9-fold cross-validation[57–59] scheme without preparing a holdout test dataset, and the results of validation were used for the hyper-parameter determination. Second, we set aside the holdout test dataset that included 11 phosphors, and an 8-fold cross-validation was implemented for the rest of the data, which included 80 phosphors, and we tested the fully trained model using the holdout test dataset. We used the optimal hyper-parameters obtained from the preceding 9-fold cross-validation process for the ensuing 8-fold cross-validation and test processes. Additionally, a leave-one-out cross-validation[57–59] was also incorporated for the 91-phosphor dataset with no holdout test dataset.

We had a similar goodness of fit for the validation irrespective of the data splitting option, *viz.*, the mean square error (MSE) and coefficient of determination (R^2) for the 9- and 8-fold cross validation and the leave-one-out cross validation, although the validations MSE and R^2 were slightly worse than those for the training for all the data-splitting schemes. Since the holdout dataset test results were similar to the 9- and 8-fold cross-validation results, the holdout dataset test results along with the 8-fold cross-validation results were accepted as the baseline in the present investigation.

Hyper-parameter optimization has been of particular concern in recent ML approaches, and the most promising strategy is known to be the use of Bayesian optimization.[55] Unlike typical deep-learning cases, however, the present problems of a small model size and a small dataset did not require such an additional optimization algorithm. We designed an allowable hyper-parameter mesh (search space) for an algorithm. Each mesh involved, at best, around 100 hyper-parameter sets (the maximum number of hyper-parameter sets was 144 for a gradient boost algorithm). We screened all the hyper-parameter sets in terms of the MSE and R^2 from the 9-fold cross-validation, and eventually pinpointed the best hyper-parameter set. All the hyper-parameter sets we tried are given in Table S1,† and the finally selected hyper-parameter set for each algorithm is highlighted in Table S1.† Details of the validation MSE and R^2 values for all the hyper-parameter sets are also listed in Table S2.†

Results and discussion

Data acquisition and descriptor extraction

Although a much greater number of Eu^{2+}-activated phosphors have been reported thus far, we incorporated 91 Eu^{2+}-activated phosphors, the chemical formulae of which are listed up in

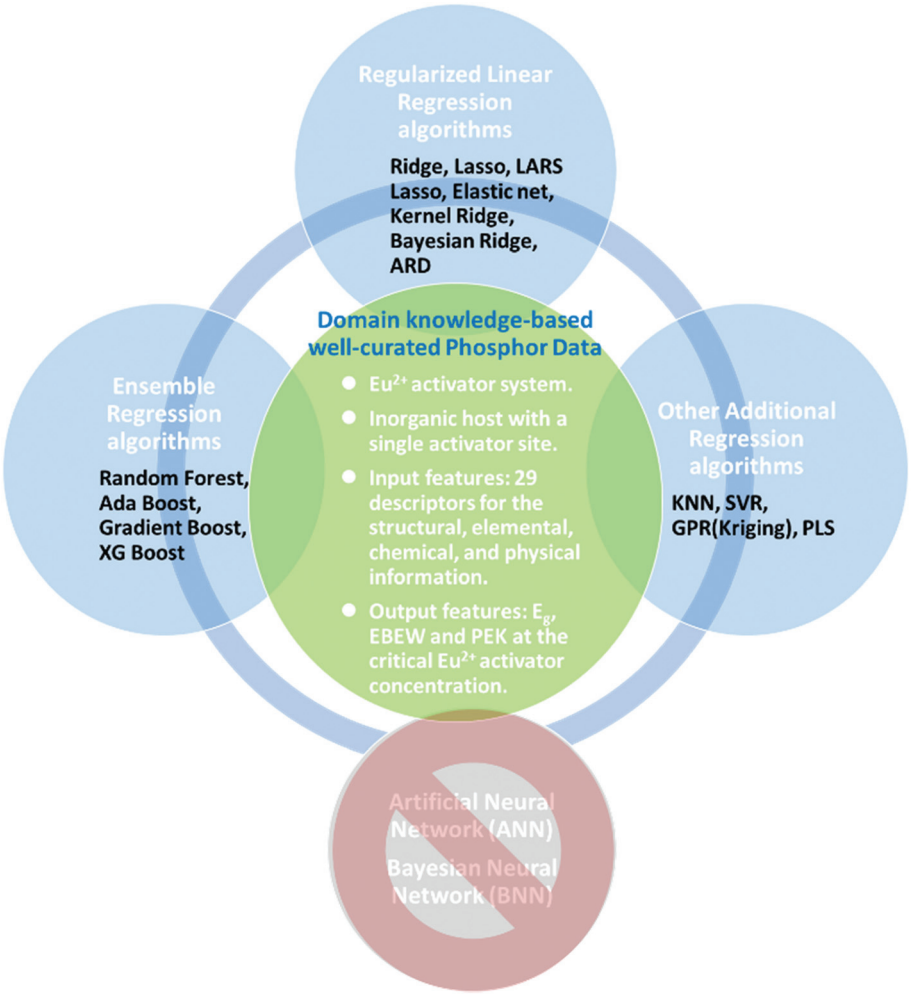


Fig. 1 The schematic representation of all procedures for the synthetic XRD data preparation. The prohibition mark implies that neither ANN nor BNN worked out for our phosphor dataset.

Table S3.† In addition, all the reference papers, from which the 91 entries came from, are listed up in the ESI.† The peak emission wavelength (PEW) that was selected as one of output features is highly dependent on many extrinsic factors such as inhomogeneous peak broadening, activator concentration, and site occupancy. The excitation band edge wavelength (EBEW) is also influenced by these external factors, although it is not as severe as the PEW. Therefore, our approach was restricted to Eu^{2+} activated phosphors that have only a single Wyckoff site for the Eu^{2+} activator to minimize the extrinsic influence. In addition, although the conditions for a single Wyckoff site were met, we excluded all the binary host compounds since every commercially available LED phosphor has multi-element hosts (ternary or higher), and thereby the descriptor extraction was devised more favorably for a multi-element host. More importantly, the data acquisition was restricted to examples wherein so-called concentration-quenching data were available. It should be noted that the most influential extrinsic variable affecting the PEW is the Eu^{2+} activator concentration. The PEW at the critical Eu^{2+} activator concentration is the only variable that could consistently be predicted using the material descriptors that we employed. Here, the critical Eu^{2+} activator concentration (x_c) corresponds to a concentration that exhibits the highest PL intensity. At an arbitrary Eu^{2+} activator concentration, the PEW would be a meaningless random variable.

The PEW data is lacking because it was extracted from a limited number of reports that provided us with clear concentration-quenching data when the single Wyckoff site conditions were met. Although EBEW seems less affected by the Eu^{2+} activator concentration, it is recommended that the EBEW be collected at the critical Eu^{2+} activator concentration. As a consequence, we were able to secure only 91 Eu^{2+}-activated phosphors that met the above-described requirement. A scientifically reasonable PEW would be the so-called zero phonon line (ZPL) that could be measured at a cryogenic temperature for an extremely dilute Eu^{2+} activator concentration. The ZPL indicates the unification between EBEW and PEW. Using the ZPL data would seem to make the present ML approach much more robust, but this would make no sense

from a practical point of view since no such data are available in the field of real-world phosphor research.

The E_g data were collected through two routes; we collected E_g values for 40 entries that are available in the materials project database,[17] and we DFT-calculated the other 51 entries in the present investigation using the same calculation conditions that are used for the materials project database. It should be noted that all the calculated E_g data were based on a PBE-GGA exchange correlation functional, so that they are underestimated below the true band gap. However, it would be no problem to use these GGA-calculated E_g values since it is well-known that there is a linear relationship between GGA-calculated and real E_g values.[60–62]

In order to systematically extract descriptors (input features), the Eu^{2+}-activated phosphors take the form prescribed in the ANX formula $A_aB_bC_cX_x$:Eu^{2+}. In the ANX formula, A, B, and C denote cation sites and X denotes an anion site, each of which has an independent Wyckoff site. The small letters (a, b, c, and x), the suffixes of the cations and the anion in the ANX formula, denote the respective stoichiometries. Activator ions tend to occupy the A site, which normally consists of alkali-earth-elements. The B site normally is occupied by alkali-earth-elements or lanthanides and is a non-activator site, while the C site is a networking element that is the most important in determining the entire structure network of a host compound. The C site consists of light elements such as B, Al, Si, or P, which constitutes borates, aluminates, silicates or phosphates, and sometimes Li, Sc, or Mg are also present at the C site. The C site forms either a tetrahedron or an octahedron with neighboring anions *via* bridging or triple points and creates a two- or three-dimensional network to provide the overall structural framework of the host compound. It should be noted that the stoichiometry (a, b, c, and x) of the host compound are not considered descriptors since they are implicitly involved in other structural descriptors.

The structural descriptors define the simplified local polyhedral information around the activator sites. At these sites the activator-anion ligand polyhedron and the activator-cation polyhedron are parameterized as descriptors. That means that on the top of the A–X polyhedron, the activator-cation polyhedron is taken as descriptors such as A–A, A–B, and A–C polyhedra consisting of the nearest-neighboring cations around the A site. The coordination number and the average distance of every polyhedron around the activator site are defined as descriptors such as $N_{A–X}$, $N_{A–A}$, $N_{A–B}$, and $N_{A–C}$ for the coordination number, and $d_{A–X}$, $d_{A–A}$, $d_{A–B}$, and $d_{A–C}$ for the average distance. Further, in the ML process, in order to reasonably account for those phosphors that have no B site in the structure, a reciprocal value of the distance was used such that $1/d_{A–B}$ becomes zero for a nonexistent B site. The local structures around the activator site that are used for the descriptor extraction could be considered an alternative interpretation of the entire structure.

The local structure around the Eu^{2+}activator involves the nearest neighbors such as X, A, B, and C and are represented as A–X, A–A, A–B, and A–C for the 91 different phosphor hosts, as shown in Fig. 2. The criteria adopted for deciding the nearest neighbor was based on the first substantial rise in the magnitude of interatomic distance. However, in certain cases,

(a)

	chemical formula	A-X	A-A	A-B	A-C		chemical formula	A-X	A-A	A-B	A-C
1	$Na_2BaSi_2O_6$					24	$SrSi_6N_8$				
2	$BaSi_7N_{10}$					25	$BaAlSi_5O_2N_7$				
3	$SrSiN_2$					26	$KMg_4(PO_4)_3$				
4	$MgAl(PO_4)O$					27	$BaSi_2O_2N_2$				
5	$SrZnP_2O_7$					28	$MgSO_4$				
6	$BaAl_2Si_2O_8$					29	$CaSO_4$				
7	$BaAl_2Si_2O_8$					30	$BaSiN_2$				
8	$CaAl_4O_7$					31	$RbBa(PO_4)$				
9	$Ba_2ZnSi_2O_7$					32	$BaSO_4$				
10	$CaMgSi_2O_6$					33	$SrB_2Si_2O_8$				
11	$BaMg_2Si_2O_7$					34	$BaB_2Si_2O_8$				
12	$Ba_2MgSi_2O_7$					35	$SrSO_4$				
13	$SrSiAl_2O_3N_2$					36	$Sr[Mg_2Al_2N_4]$				
14	$BaSiAl_2O_3N_2$					37	$Sr_8(Si_4O_{12})Cl_8$				
15	Li_2BaSiO_4					38	$Ca[LiAl_3N_4]$				
16	$LiSiON$					39	$Sr[Mg_3SiN_4]$				
17	$CaAlSiN_3$					40	BaB_8O_{13}				
18	$BaZn_2Si_2O_7$					41	$Ca_2Al_2SiO_7$ $(Ca_2Al_{1.77}Si_{1.23}O_{6.77}N_{0.23})$				
19	$BaAlSi_4O_3N_3$					42	$SrCaSiAl_2O_7$				
20	$SrAlSiN_3$					43	$Sr_2Al_{1.6}Si_{1.4}O_{6.6}N_{0.4}$				
21	$MgAlSiN_3$					44	$Sr_2ZnSi_2O_7$				
22	$Sr(Al_{0.3}Si_{0.7})_4(N_{0.8}O_{0.2})_6$					45	$Ca_2MgSi_2O_7$				
23	$BaSi_6N_8O$					46	$Sr_2MgSi_2O_7$				

Li Be B N O F Na Mg Al Si P S K Ca Sc Zn Ga Rb Sr Y Zr Cd Cs Ba Gd Hf

(b)

	chemical formula	A-X	A-A	A-B	A-C		chemical formula	A-X	A-A	A-B	A-C
47	$SrLaGa_3S_6O$					70	$Sr_8Al_{12}O_{24}S_2$				
48	Li_2CaSiO_4					71	$Sr_3Bi(PO_4)_3$				
49	$KSrBP_2O_8$					72	$Sr_{12}Al_{14}O_{32}Cl_2$				
50	$KBaBP_2O_8$					73	$Ca_{12}Al_{14}O_{32}Cl_2$				
51	$BaMgSi_4O_{10}$					74	$CsAlSi_2O_6$				
52	$Sr_3GdNa(PO_4)_3F$					75	$Ba_2LiB_5O_{10}$				
53	$BaAl_2Si_2O_8$					76	$BaAl_2B_2O_7$				
54	$SrSi_9Al_{19}ON_{31}$					77	$BaBeSiO_4$				
55	Li_2SrSiO_4					78	$BaGa_2SiS_6$				
56	$SrBPO_5$					79	$BaMg_2SiN_4$				
57	Ca-a-SiAlON $(Ca_{0.375}Si_{11.25}Al_{0.75}N_{16})$					80	$BaMgP_2O_7$				
58	Sr-a-SiAlON $(Sr_{0.375}Si_{11.25}Al_{0.75}N_{16})$					81	$CsMgPO_4$				
59	$LiCaPO_4$					82	$LiSrPO_4$				
60	$BaAlBO_3F_2$					83	$RbLi(Li_3SiO_4)_2$				
61	$CaZrBAl_9O_{18}$					84	$Sr_2Al_2SiO_7$				
62	$SrYSi_4N_7$					85	$Sr_4LiAl_{11}N_{14}$				
63	$CaScSi_4N_7$					86	$SrAl_4O_7$				
64	$SrScSi_4N_7$					87	SrB_4O_7				
65	$BaScSi_4N_7$					88	$BaSi_2O_5$				
66	$BaZrSi_3O_9$					89	$CdSrP_2O_7$				
67	$BaHfSi_3O_9$					90	$Sr_3NaY(PO_4)_3F$				
68	$Mg_2Al_4Si_5O_{18}$					91	$SrMgAl_{10}O_{17}$				
69	Sr_2ScAlO_5										

Li Be B N O F Na Mg Al Si P S K Ca Sc Zn Ga Rb Sr Y Zr Cd Cs Ba Gd Hf

Fig. 2 Schematics for A–X, A–A, A–B, and A–C local structures for 91 Eu^{2+}-activated phosphors. Atoms are represented by different colors, as shown below. Also, the relative distance obeys the actual length scale. The number represents corresponding phosphors listed in Table S3 in the ESI.†

where the interatomic distance continuously increases rather than showing a step-rising trend, the nearest neighbors were consistently decided at a distance change of 10%, although Pan *et al.* have reported a smart coordination decision protocol.[63] Local structures with A–A and A–B coordination sometimes do not constitute a polyhedron but simply result in 1-or 2-dimensional shapes. This type of coordination is equally important in deciding the optical properties since it serves as the routes for inter-activator energy transfer. In traditional approaches to studying luminescence properties by focusing on either isolated or extremely dilute activator systems, the activator-anion local structure based on the framework of the crystal field strength and the nephelauxetic effect is a major concern, while the cation–cation local structure around the activator site is out of concern. It is noteworthy, however, that the cation neighbors around an activator also have a significant impact on luminescence properties (PEW in particular) when real-world phosphors with a practical-activator concentration are a concern. The energy-transfer mechanism matters in real-world phosphors, wherein the A–A, A–B, and A–C variables play a more influential role than the A–X variable for predicting the PEW.

In addition to the above-described structural descriptors (N_{A-X}, N_{A-A}, N_{A-B}, N_{A-C}, d_{A-X}, d_{A-A}, d_{A-B}, and d_{A-C}) that designate the local structure around the activator site, we also defined another 12 descriptors that indicate other elemental information of the constituent elements occupying the A, B, C, and X sites. These descriptors are the atomic number (Z_A, Z_B, Z_C, and Z_X), the electronegativity (E_A, E_B, E_C, and E_X) on the Pauling scale,[64] and the Shannon radius (R_A, R_B, R_C, and R_X) of every constituent element in the respective local environments and in the respective valence states.[65] Special consideration was given to cases where the A, B, C, and X sites were occupied by more than one element. In this case, weighted average values were obtained according to the elemental fraction in order to evaluate the elemental characteristic parameters. In addition, lattice parameter anisotropy, lattice angles, lattice volume, and theoretical density were also adopted as descriptors (a/c, b/c, β, γ, V and ρ). Here, the lattice parameter was set in such a way that $c \geq b \geq a$. Angle α was omitted from the descriptors because our dataset contained no triclinic structure. Finally, basic symmetry descriptors such as the space group number (SG), the activator site symmetry number (SS), and the activator site multiplicity (AM) were adopted. Table S3† shows the chosen 29 descriptors and their evaluation results for 91 different Eu^{2+}-activated, single-A-site phosphors.

The most interesting target variables, *viz.*, the EBEW and PEW, were adopted in units of eV at the critical Eu^{2+} activator concentration (x_c). In particular, the PEW should never be regarded as a material's intrinsic property because it is known to vary dramatically with the concentration of the Eu^{2+} activator. It is, therefore, important to have concentration-quenching data such as the emission spectra monitored as a function of the Eu^{2+} activator concentration for a given phosphor. The concentration-quenching data, however, were available for only 77 out of 91 phosphors. In the absence of concentration-quenching data, we contacted the authors of the literature presenting no concentration-quenching data to verify that the emission spectrum came from x_c.

From a strict theoretical point of view, an energy value corresponding to ZPL collected for extremely diluted model phosphors at cryogenic temperatures would be the best target variable to be predicted from the descriptors introduced above. It is, however, impossible to obtain a ZPL value for each phosphor. Further, if the ZPL data were available, the data-driven ML approach might not be that important in the field, because a theoretical approach would have been sufficient to a large extent. The ZPL data could be more attractive when considering the theoretical modeling, but those data are extremely scarce, and the conventional excitation and emission data measured at room temperature for a conventional activator concentration range is of more practical interest in the field. In addition, a clear interpretation of these data based on the theoretical approach alone would be difficult, since there are many extrinsic issues such as site occupation complicacy (involving inhomogeneous broadening), inter-ionic energy transfer, lattice phonon interaction, powder characteristics, *etc*. Therefore, a data-driven approach would be more suitable for predicting the emission energy of conventional real-world phosphors compared with using a theoretical approach.

Regression results

Since the most important issue in the present investigation is the PEW prediction from a practical point of view, we prioritized it and relegated the EBEW and E_g predictions to the level of auxiliary tasks that support the PEW prediction. The same training procedures and hold-out test dataset were applied to the PEW, EBEW, and E_g prediction models, and the results from the EBEW and E_g prediction models are summarized in the ESI (Table S4a and S4b†), while the PEW prediction results appear in Table 1. It is worth noting that the EBEW and E_g predictions exhibited a better regression fitting quality by providing a better goodness of fit by comparison with the PEW prediction. In particular, the EBEW prediction model conspicuously outperformed the others. This finding implies that the adopted input feature (descriptor) setting as well as the ML algorithm selection was validated by multiple output features such as EBEW and E_g, and the adopted descriptors still proved to be suitable for the PEW prediction.

The E_g prediction models were not of great concern in the present investigation since E_g is not a key factor affecting EBEW and PEW as far as it guarantees an insulating level that is greater than the 4f–5d level of energy. It is, however, interesting to see a good fitting quality for several E_g prediction models. The E_g prediction results could be regarded as just a successful ML example that reconfirms the validity of adopted descriptors and ML algorithms. It should be noted that the PEW prediction results presented in the main text are the baseline that exhibits the worst predictability by comparison with the EBEW and E_g prediction results, although the overall fitting quality for the PEW prediction was still acceptable in

Table 1 The training, validation and hold-out dataset test results for PEW prediction model in terms of MSE and R^2 for two data-splitting schemes: 9-cross-validation and 8-fold cross-validation with a holdout test dataset

PEW										
	9-Fold cross validation				8-Fold cross validation with hold-out dataset test					
ML algorithm	MSE (training)	R^2 (training)	MSE (validation)	R^2 (validation)	MSE (training)	R^2 (training)	MSE (validation)	R^2 (validation)	MSE (test)	R^2 (test)
Basic linear	0.011	0.80	0.026	0.48	0.011	0.80	0.028	0.41	0.033	0.38
Ridge	0.013	0.76	0.021	0.62	0.014	0.75	0.023	0.59	0.021	0.61
Lasso	0.013	0.76	0.022	0.60	0.013	0.76	0.024	0.57	0.019	0.62
LARS	0.015	0.73	0.021	0.62	0.015	0.74	0.022	0.57	0.021	0.61
Elastic net	0.013	0.77	0.022	0.61	0.013	0.77	0.025	0.51	0.018	0.66
KRR	0.016	0.71	0.021	0.62	0.017	0.70	0.022	0.59	0.019	0.64
BRR	0.013	0.77	0.022	0.61	0.014	0.76	0.023	0.58	0.021	0.61
ARD	0.014	0.75	0.021	0.61	0.014	0.75	0.025	0.55	0.020	0.62
Random forest	0.004	0.93	0.025	0.58	0.004	0.93	0.026	0.53	0.024	0.54
Ada boost	0.012	0.79	0.023	0.56	0.012	0.78	0.029	0.43	0.020	0.62
Gradient boost	0.005	0.92	0.023	0.59	0.004	0.93	0.027	0.48	0.028	0.47
XG boost	0.001	0.98	0.027	0.53	0.001	0.99	0.025	0.51	0.029	0.45
SVR	0.013	0.77	0.020	0.63	0.013	0.76	0.024	0.57	0.018	0.67
KNN	0.000	1.00	0.027	0.48	0.000	1.00	0.030	0.45	0.036	0.32
PLS	0.026	0.54	0.031	0.40	0.026	0.53	0.032	0.42	0.032	0.39
GPR	0.000	1.00	0.023	0.62	0.000	1.00	0.021	0.61	0.025	0.52

view of recent field standards. Averages for the MSE, R^2, and overfitting index for 16 ML algorithms for the hold-out test dataset and also for the validation dataset were ranked such that EBEW > E_g > FEW (the symbol '>' designates 'is better than'). On these grounds, the number of ML algorithms with an R^2 > 0.6, MSE < 0.02, and test_R^2/training_R^2 > 0.77 for the hold-out test dataset followed the same rank, *i.e.*, seven for EBEW, six for E_g, and four for FEW. Better results for the EBEW and E_g predictions appear in the ESI, as shown in Table S4 and Fig. S1.† The ML algorithms nominated for R^2 > 0.6, MSE < 0.02, and test_R^2/training_R^2 > 0.77 and the corresponding PEW, EBEW, and E_g prediction results are marked by red boxes in Fig. 4.

Table 1 shows the MSE and R^2 for PEW prediction models, which was evaluated for the training, validation, and holdout dataset test for the 9- and 8-fold cross-validations along with the holdout dataset test. Fig. 3a and b graphically shows the same results. The goodness of fit remained almost identical for both the 9- and 8-fold cross-validations as shown in Table 1 and Fig. 3a and b, and the leave-one-out cross-validation results were also similar, which is available in the ESI (Table S5 and Fig. S3†). Despite the superior fitting quality for training (*i.e.*, lower MSE and higher R^2 for training), we placed greater emphasis on the validation MSE and R^2, and eventually much more on the holdout dataset test. While the overall MSE level was approximately 10^{-3}–10^{-2} for training, the validation MSE increased slightly and the holdout dataset test MSE results were similar to the validation results. The overall R^2 level for the training was 0.7–1, and the R^2 level for both the validation and the holdout dataset test was approximately 0.6. If R^2 for the validation (or the holdout dataset test) exceeds 0.5, then the regression results are generally acceptable by the statistics research society,[66] although the conventional standard for the R^2 level has not been clearly defined yet. It should be noted that training the MSE and R^2 for well-known parameter-free regression algorithms such as KNN and GPR reached the perfect level, *i.e.*, 0 and 1, respectively. This was due to the inherent non-parametric trait of the KNN and GPR algorithms. The basic linear, KNN, and PLS algorithms led to unacceptable validation and test results, while the others exhibited almost similar levels of validation and test results for MSE and R^2. As evidenced by the EBEW and E_g prediction results shown in Table S4 and Fig. S1,† the overall fitting quality for the EBEW and E_g prediction is better than that for the PEW prediction, namely, the overall levels of MSE (and R^2) were lower (and higher) than those for the PEW prediction.

Rather than the absolute value level for MSE and R^2, the over-fitting (high variance) problem would be much more important in judging the regression fitting quality in the case of an extremely small dataset, as with the present case. The ratio between the training and validation MSE (and R^2) is indicative of the level of over-fitting, which is referred to as the 'over-fitting index'. Fig. 3c and Table S6a† shows the over-fitting index for the PEW prediction in the range of 0–1, which is defined as training_MSE/validation_MSE and validation_R^2/training_R^2. A higher over-fitting index indicates a better fit, *i.e.*, 1 is reached in an ideal case. Fig. 3d also shows the over-fitting index for the PEW prediction, defined as training_MSE/test_MSE and test_R^2/training_R^2, which are similar to the training_MSE/validation_MSE and validation_R^2/training_R^2. It is evident that the basic linear regression gave rise to an atrocious over-fitting. In addition, both the KNN and GPR also exhibited extremely low over-fitting index values, *viz.* zero for both the training_MSE/validation_MSE and training_MSE/

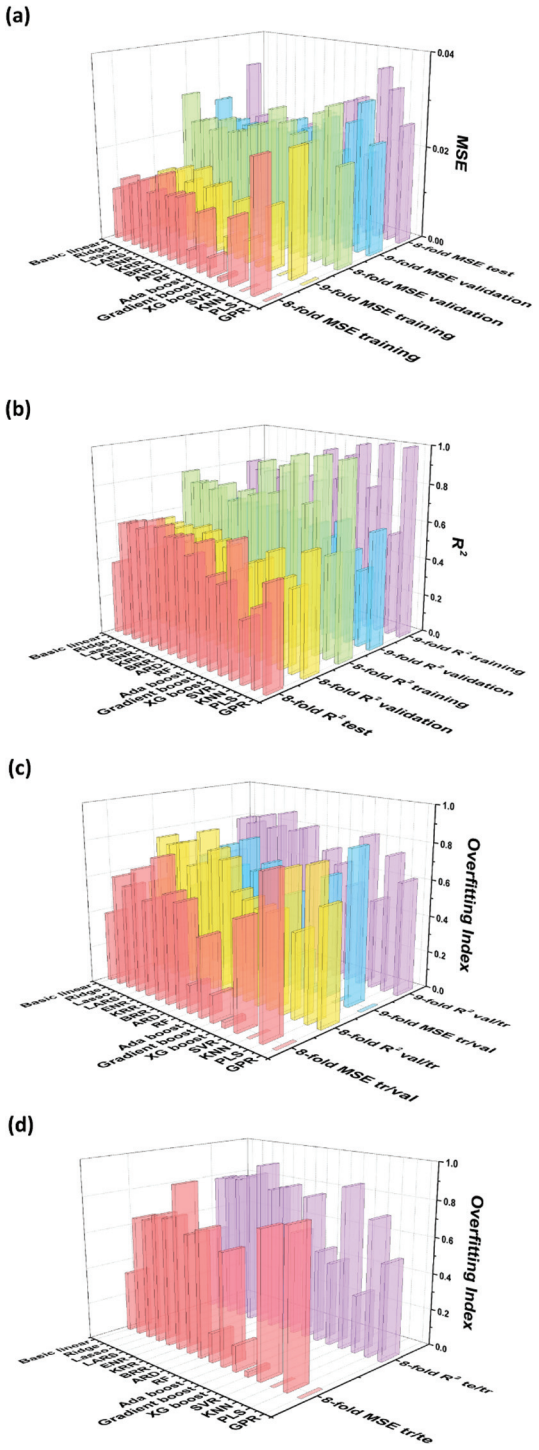


Fig. 3 The training, validation and hold-out dataset test results for PEW prediction in terms of (a) MSE and (b) R^2 for 9 cross validation and 8-fold cross validation with a holdout dataset test, (c) the over-fitting index defined as training_MSE/validation_MSE and validation_R^2/training_R^2, and (d) the over-fitting index defined as training_MSE/test_MSE and test_R^2/training_R^2. Each over-fitting index value is listed up in Table S6.†

test_MSE. The KNN and GPR gave rise to a serious over-fitting problem, as in the case of the basic linear regression, because the KNN and GPR algorithms produced a perfect fit for a training dataset due to their intrinsic non-parametric traits, as already discussed above. Thus, it would be irrational to equate these low overfitting index values to those of the heavily over-fitted basic linear regression, particularly in the case of GPR that gave acceptable validation/test MSE and R^2 values. Accordingly, we regarded GPR as an acceptable ML algorithm regardless of the over-fitting index value. The ensemble algorithms such as RF, Ada boost, gradient boost, and XG boost also raised over-fitting issues. On the other hand, it is evident that a certain degree of regularization took place for the other regularization-involved linear regression algorithms. The over-fitting problem was also considerably improved in the EBEW and E_g prediction, as shown in Fig. S1(c), S1(d), S1(g), and S1(h) and Table S6b and S6c in the ESI.†

It should be noted that regularization-involved ML algorithms were introduced since there was a severe training data shortage problem in the present investigation. The regularization-involved linear regression algorithms outperformed the ensemble algorithms, and the SVR also gave an acceptable over-fitting index, as shown in Fig. 3 and Table S6a.† When comparing the basic linear regression and the other regularization-involved linear regression algorithms, it is apparent that the validation (and test) MSE and R^2 were improved at the expense of the training MSE and R^2 for the regularization-involved linear regression algorithms. Fig. 4a and S2a† shows plots of the predicted *vs.* experimental emission energy, and the training dataset led to a relatively good fit for the basic linear regression (upper-leftmost corner), but the validation and test datasets gave slightly worse fits by comparison with those of the other regularization-involved linear regression algorithms.

The same trend was observed for EBEW and E_g prediction models (Table S6b and S6c,† Fig. 4b, c, and Fig. S1, S2b, S2c†), although the overall level of overfitting index was higher than that for the PEW prediction model.

In addition, the ANN (or DNN) results were omitted from Table 1, since the regression results (over-fitting in particular) were even worse than any of the other algorithms listed in Table 1. A paucity of data never allows for an ANN model since even architecture with a single hidden layer involves too many parameters by comparison with the training dataset size, and the Bayesian neural network[67] was not viable for the same reason. Accordingly, the basic linear regression was taken as our baseline result. With the noted exceptions of ANN, basic linear, KNN, and PLS, all the other algorithms involve a certain degree of regularization functions. Consequently, the ANN regression gave rise to the poorest regression results, and thereafter the basic linear regression followed. The regularization-involved linear regression algorithms such as ridge, LASSO, elastic net, KRR, and BRR mitigated the problem of over-fitting to a certain extent, and the absolute MSE and R^2 levels were also acceptable, as shown in Table 1. The ARD and LARS results were slightly deteriorated, but KNN and PLS gave

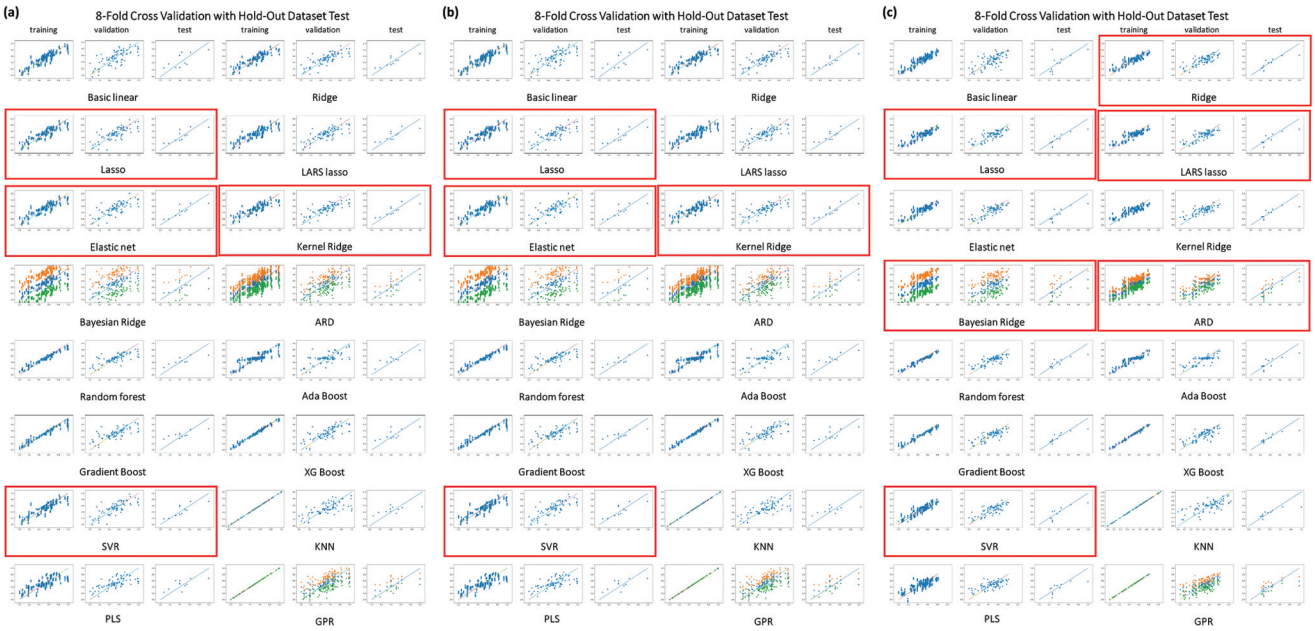


Fig. 4 Plots of predicted *vs.* experimental plots of predicted *vs.* experimental (a) PEW, (b) EBEW, and (c) E_g for training, validation, and hold-out test datasets for 8-fold cross validation.

an unacceptable regression quality. Consequently, we nominated four promising ML algorithms; LASSO, elastic net, KRR, and SVR for the PEW prediction. These ML algorithms met the condition $R^2 > 0.6$, MSE < 0.02, and overfitting index (test_R^2/training_R^2) > 0.77 for the hold-out dataset test. It is noted, however, that all the regularization-involved linear regression algorithms gave a fitting quality that was as good as those nominated for $R^2 > 0.6$, MSE < 0.02, and test_R^2/training_R^2 > 0.77. The same trend was also observed for the EBEW and E_g predictions. Consequently, it was revealed that only the regularization-involved linear regression algorithms are suitable for the present ML approach when there is a dearth of data.

As shown in Fig. 4 and S2,† the BRR, ARD and GPR regression results exhibited conspicuous and distinctive fitting. These three algorithms are all based on the Bayesian approach. Bayesian approach-involved algorithms give a range of confidence around the predicted mean rather than a deterministic prediction. The amber and green dots for the BRR, ARD and GPR results designate the standard deviation range, as shown in Fig. 4 and S2.† These sorts of Bayesian approaches would be more desirable than the other customary regularization strategies due to the fact that uncertainty in the prediction can also be formulated.

The main focus of the present investigation was to obtain a plausible regression result for real-world dirty datasets that are far smaller than what the conventional ANN approach requires. In fact, the over-fitting problem was unavoidable, but we were able to mitigate it to a certain extent by introducing regularization-involved algorithms and thereby we secured generally acceptable regression results, albeit below the level of conventional deep learning. Although it is practically impossible to collect a sufficient level of clean data from either industry or academia, the ML approach merits an application to phosphor research. The so-called integrated ML platform that we proposed in the present investigation could be a tentative solution for both the problems of the data paucity and the dirty data. Of course, the use of an ML approach based on a sufficient amount of the synthetic data that is available in well-known DFT-driven databases[17,68–70] would definitely return quite an excellent regression result. On the other hand, we only focused on real-world data of a dirty nature in the present investigation, and the worst problem was that the paucity of data led to serious over-fitting (high valiance) that could not be completely sorted out.

The meaning of the term 'dirty' is two-fold. The first aspect concerns the experimental inaccuracies (or inconsistency) originating from many different material syntheses and characterization platforms. Experimentally evaluated lattice parameters that greatly affect the descriptors were acquired from various research groups, so that they inherently involved a certain degree of errors. The same complication applies to the PEW and, EBEW measurement. The second aspect of the term 'dirty' is a data distribution-related problem. When the collected data are not identically and independently distributed (IID) random data, the distribution for some descriptors is discrete and biased. The input-feature (descriptor) distribution does not necessarily have to be an IID-Gaussian distribution as far as the output loss (the difference between real and model-predicted outputs) is approximated to a Gaussian. However, such a highly biased non-IID data distribution would not be beneficial to ML-based regression. Fig. 5a shows 1-D data distribution for every input/output feature in the histogram

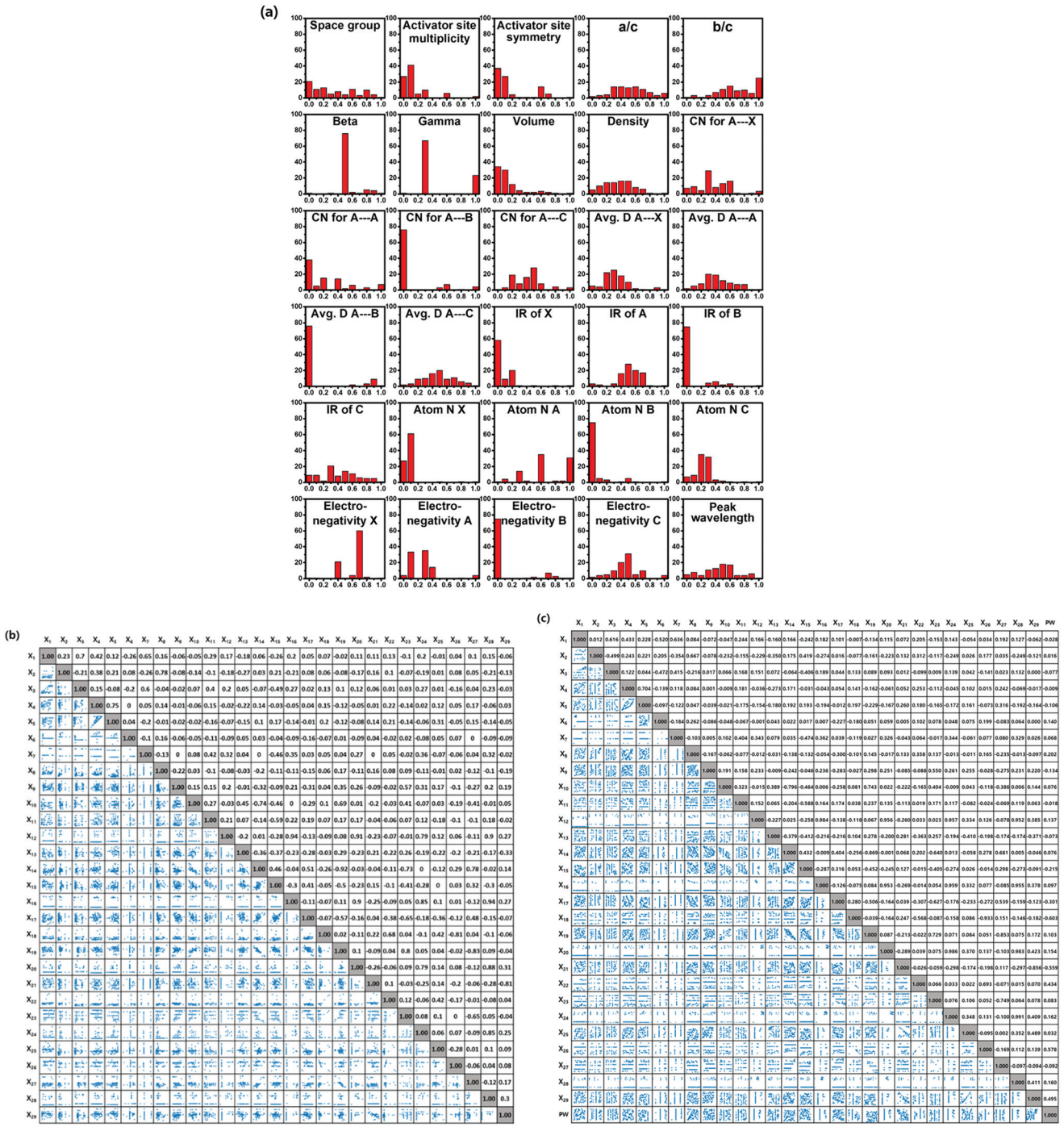


Fig. 5 (a) 1-D data distribution for each of the 29-input and 1-output features, (b) the Pearson correlation coefficient matrix for 29-input features; the upper off-diagonal components are the Pearson correlation coefficients, and the lower off-diagonal components are pair-wise 2-D data distribution plots. (c) Spearman correlation coefficient matrix for 29-input features. The lower off-diagonal components are a pair-wise 2-D distribution of the data rank.

format, and Fig. 5b exhibits a more convenient representation of data distribution, that is, the pair-wise 2-D plots along with Pearson-correlation coefficient matrix. In addition, the Spearman correlation coefficient matrix is also given in Fig. 5c. The Spearman correlation coefficient is the Pearson correlation coefficient for ranks, which can take care of a non-linear relationship and also rule out the outlier effect.[71] Accordingly, the pair-wise 2-D plots in Fig. 5c are a data rank distribution rather than a raw data distribution. There is no notable difference between Pearson and Spearman correlation coefficient matrices and both these forms of correlation data equally exhibit a highly biased non-IID data nature.

If data were collected from a uniform data production platform and the amount was significantly larger, then the MSE and R^2 levels for the validation (or test) would reach that of typical synthetic data from deep learning. The goal of the present investigation, however, was to run ML algorithms for dirty real-world data and even for a dearth of data, which we are faced with in the phosphor research society. We deal with neither IID data nor so-called big data (a large-scale dataset). Big IID data for ML approaches to materials science is tangible in only synthetic forms. A uniform high throughput experimental data production platform, *e.g.*, data collected from a single lab using the same apparatuses, would rule out the dirty nature of data and thus give rise to far better regression results. The data used here, however, were collected from a variety of literature sources, which could never constitute a IID random dataset originating from a certain expert intervention during a data-production procedure.

Our dataset involved many outliers due to inconsistent experiments. Nonetheless, the proposed 'integrated ML platform' deserves application to such a dirty engineering dataset, and we determined the best way to treat such data.

$$D = \varepsilon_c + \frac{\varepsilon_{cfs}}{F} - \varepsilon_s(\text{free})$$

Theoretical interpretation for PEW and EBEW through its surrogate ML model

Ignoring all the extrinsic effects on the excitation and emission energy (EBEW and PEW), a theoretical (or semi-empirical) interpretation of the excitation and emission energy is possible, as reported by Dorenbos.[72–75] Fig. 6 shows the energy diagram for the 5d energy levels of Ce^{3+} and Eu^{2+} activators in a certain host. The excitation and emission energy can be inferred from the free-ion state of energy by inferring a centroid shift (ε_c), a total shift (D), crystal field splitting (ε_{cfs}), and a Stokes shift (ΔS). ε_s(free) is the energy difference between centroid position and the lowest 5d level of the free Eu^{2+} ion, which can be ignored since it is very small by comparison with D and ε_{cfs}.

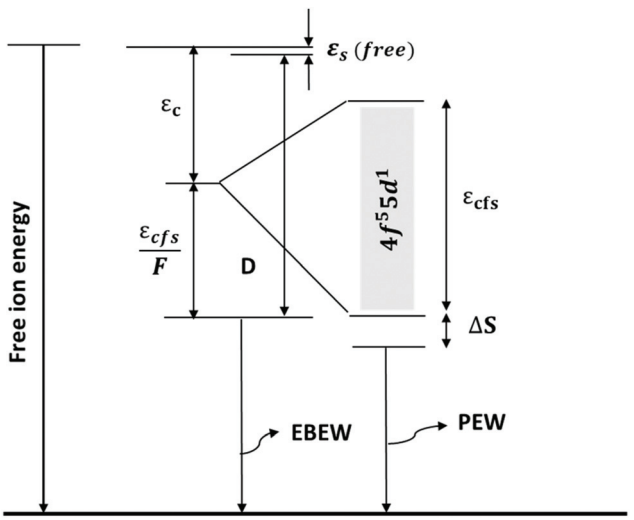


Fig. 6 The schematics for 5d energy level and theoretical models elucidating the total shift (all the variables and parameters appearing in the model are accounted for in the manuscript).

The theoretical (or semi-empirical) model[72–75] describes ε_c and ε_{cfs} in terms of several basic measurable variables such as activator-anion ligand distance (R_i), coordination number (N), anion polarizability (α_{sp}), ionic size difference (ΔR), and the ratios of crystal field splitting (F). This theoretical model[72–75] originally describes Ce^{3+} but it also holds for Eu^{2+} since it is well-known that the total shift, the Stokes shift, the centroid shift and the total crystal field splitting of the 5d levels of Eu^{2+} and Ce^{3+} all are linearly related to one another.[75]

Regardless of whether the ligand polarization model[76,77] or the covalency model[77,78] was adopted, the centroid shift (ε_c) and the crystal field splitting (ε_{cfs}) have been interpreted in terms of the activator-ligand local structure that is traditionally parameterized as the activator-anion ligand distance and coordination number. Namely, ε_c and ε_{cfs} can be interpreted using the A–X local environment. Once the total shift (D) was evaluated from the local structure-based theoretical (or semi-empirical) models for ε_c and ε_{cfs}, prediction, the EBEW can be exactly estimated since the EBEW is the difference between the free ion energy and the total shift (D), as shown by the energy diagram in Fig. 6.

Since the emission energy should be greatly affected by the total shift and the crystal splitting, the emission energy could be also a function of the above-described variables that designate the activator-anion ligand (A–X) local structure and the trait of the constituent elements for the A–X polyhedron. In contrast to the A–X local environment-related ε_c and ε_{cfs}, the emission energy is greatly affected by the Stokes shift. The Stokes shift significantly differs from one host to another. According to the Franck–Condon theory,[79] the Stokes shift seems to be closely related to the host lattice stiffness and to multi-phonon behaviors when the configuration coordinate model is accounted for. The Stokes shift is not simply interpreted by the above-mentioned A–X local environment variables only. Although the EBEW can be directly evaluated from the theoretical model for the total shift prediction, no theoretical model for the emission energy prediction was available due the Stokes shift complication. The Stokes shift can be ignored if the zero phonon energy is available for an extremely dilute system at a cryogenic temperature, but it is impractical in real-world phosphor research.

The theoretical model for the prediction of the total shift (*i.e.*, EBEW) deserved to be tested using the collected data. A direct application of the theoretical model, however, was practically impossible since we could scarcely collect correct crystal field-splitting data. The exact evaluation of the crystal field-splitting from the conventional PLE spectrum is limited. Both α_{sp} and F were unobtainable as well. Nonetheless, we indirectly tested the theoretical model using a surrogate ML model. For this undertaking, we extracted 9 relevant descriptors out of a

total of 29. The reduced number of descriptors was either directly or implicitly related to the variables and parameters (R_i, N, α_{sp}, F, and ΔR) appearing in the theoretical model, which affected the centroid shift and crystal field splitting and led to a total shift, *i.e.*, to EBEW and finally to PEW. The extracted descriptors, which were compatible with the variables and parameters (R_i, N, α_{sp}, F, and ΔR), were the A–X distance (d_{A-X}), the activator site symmetry (SS), the Shannon ionic radius for A and X (R_A and R_X), the atomic number for A and X (Z_A and Z_X), the electronegativity for A and X (E_A and E_X), and the coordination number for A–X polyhedron (N_{A-X}). These 9 descriptors were used for the surrogate ML model. The regression results are listed for all 9-descriptor surrogate ML models in Table S7.† Fig. S4† shows the predicted PEW, EBEW, and E_g *versus* the real values for all 9-descriptor surrogate ML models. The fitting quality was deteriorated by comparison with the 29 descriptor regression results shown in Fig. 4 and S2.† When the number of descriptors was reduced dramatically from 29 to 9, the fitting quality deterioration for PEW prediction was more conspicuous than for the EBEW prediction, as evidenced in Fig. 4 and S4.† This means that the surrogate model associated with only the A–X local structure would never be viable for the prediction of EBEW and PEW, which indicates that these 9 descriptors were insufficient and a greater number of descriptors would be required to account for the excitation and emission energy. It appears more reasonable to use the sixth power of the A–X bond length (d^6_{A-X}) since the theoretical model is described as a function of R^6_{eff}.[76,77] However, the dataset including d^6_{A-X} never yielded better regression results for all the 9-input-feature surrogate ML models.

It is clear that the surrogate ML model based on a reduced number of descriptors, which was compatible with the theoretical model, never gave acceptable test results in contrast to the 29-input-feature ML models. This finding implies that only the A–X environments are insufficient to account for the excitation and emission energy (EBEW and PEW). The first reason for the inapplicability of the surrogate model, *i.e.*, the theoretical model is that the theoretical model would in principle hold only for an ideal case of very dilute activator concentration that would prevent any type of inter-activator interaction such as an energy transfer. Since the excitation and emission energy (EBEW and PEW) evaluated at a practical level of activator concentration was of concern in the present investigation, neither the theoretical model nor its equivalent surrogate ML model (9-input-feature ML model) was viable. As already discussed above, an incapable predictability for Stokes shift would be one of the major reasons for the unacceptable predictability of the 9-input-feature surrogate ML modeling for PEW prediction. The 29-input-feature system might have implicitly incorporated the Stokes shift and thereby an acceptable predictability was achieved, although we are unaware of how this functions in the ML process. In addition, PEW (or EBEW) could be shifted by inhomogeneous broadening due to the local structure fluctuation in real-world phosphors, and this is one of the reasons for the unacceptable predictability of the 9-input feature surrogate ML model that simulates a theoretical model.

We never denied the validity of the theoretical model but instead we are quite sure of its scientific validity. It should be noted, however, that the real-world data that we used here are not suitable for a theoretical model, since these data include various extrinsic traits. The application of a theoretical model is limited to an ideal case that exhibits a very homogeneous local structure, a very dilute system that guarantees no interionic interactions, and a sustained lattice vibrational system at low temperatures. This sort of ideal phosphor would never be easily found in the real world. Strictly speaking, this is not an ideal phosphor from a practical point of view and is ideal for only a theoretical model application. Although the ML model that we suggest lags far behind the theoretical approach in terms of understandable logics and scientific merit, we believe that the ML approach could outperform the theoretical model when practical problems are a real concern.

Conclusions

An integrated ML model platform involving 18 algorithms was developed to predict the peak emission wavelength (PEW), excitation band edge wavelength (EBEW), and band gap (E_g) from structural, elemental, chemical, and physical descriptors. The 91 Eu^{2+}-activated phosphors that provided the Eu^{2+}-activator with a single Wyckoff site were extracted from the literature. The PEW and EBEW data for a critical Eu^{2+} activator concentration were collected from the literature where the concentration quenching data were available.

Regularization-involved ML algorithms outperformed both the basic linear regression and ANN models. Well-known ANNs (or DNNs) were never viable due to the problem of a paucity of data. The suggested ML model platform could be a tentative ML solution to tackle the real-world problems of a dearth of data that are commonly confronted in the physical science research society. Statistically, regularization-involved linear regression algorithms seem to be the best, but it would be extremely risky to choose a single ML algorithm based only on goodness of fit and overfitting index. Due to the *ad hoc* heuristic nature of data-driven approaches, it would be inappropriate to introduce only a single ML algorithm despite it exhibited the best goodness of fit and overfitting index, but a group of ML algorithms, just like the integrated ML platform, could be a better option for a single particular problem with a single dataset. In this way we can get a reliable PEW prediction result by averaging the prediction results from four acceptable ML algorithms that meet the condition $R^2 > 0.6$, MSE < 0.02, and overfitting index (test_R^2/training_R^2) > 0.77, such as LASSO, elastic net, KRR, and SVR. Similarly, we pinpointed seven and six ML algorithms for EBEW and E_g predictions, respectively.

A well-known theoretical model wherein both centroid shift and crystal field splitting that implicitly led to EBEW and PEW predictions could be predicted from the A–X local environ-

mental information, was simulated by employing a surrogate ML model and adopting a set of appropriate descriptors that could be transformed from those appearing in the theoretical model. A surrogate ML model that was supposedly equivalent to the theoretical model did not work, by contrast to the 29-descriptor-based ML model platform that worked in a proper manner. While the theoretical model could work for certain ideal cases (*e.g.*, a very dilute activator concentration eliminating any types of interionic interactions and a very low temperature banishing the Stokes shift), the ML model platform can perform practical EBEW and PEW prediction tasks for ordinary phosphors confronted in the field of engineering.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was supported by the Creative Materials Discovery Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning (2015M3D1A1069705), (2021R1A2C1011642) and (2021R1A2C1009144), and partly by the Alchemist Project (20012196), and Digital manufacturing platform (N0002598) funded by MOTIE, Korea.

Notes and references

1 D. Durach, L. Neudert, P. J. Schmidt, O. Oeckler and W. Schnick, $La_3BaSi_5N_9O_2$:Ce^{3+}-A yellow phosphor with an unprecedented tetrahedra network structure investigated by combination of electron microscopy and synchrotron X-ray diffraction, *Chem. Mater.*, 2015, **27**, 4832–4838.

2 Q.-Q. Zhu, L. Wang, N. Hirosaki, L. Y. Hao, X. Xu and R.-J. Xie, Extra-Broad Band Orange-Emitting Ce^{3+}-Doped $Y_3Si_5N_9O$ Phosphor for Solid-State Lighting: Electronic, Crystal Structures and Luminescence Properties, *Chem. Mater.*, 2016, **28**, 4829–4839.

3 N. Hirosaki, T. Takeda, S. Funahashi and R.-J. Xie, Discovery of New Nitridosilicate Phosphors for Solid State Lighting by the Single-Particle-Diagnosis Approach, *Chem. Mater.*, 2014, **26**, 4280–4288.

4 R. Gautier, X. Li, Z. Xia and F. Massuyeau, Two-Step Design of a Single-Doped White Phosphor with High Color Rendering, *J. Am. Chem. Soc.*, 2017, **139**, 1436–1439.

5 H. Liao, M. Zhao, M. S. Molokeev, Q. Liu and Z. Xia, Learning from a Mineral Structure toward an Ultra-Narrow-Band Blue-Emitting Silicate Phosphor $RbNa_3(Li_3SiO_4)_4$:Eu^{2+}, *Angew. Chem.*, 2018, **130**, 11902–11905.

6 M.-H. Fang, C. O. M. Mariano, P.-Y. Chen, S.-F. Hu and R.-S. Liu, Cuboid-Size-Controlled Color-Tunable Eu-Doped Alkali-Lithosilicate Phosphors, *Chem. Mater.*, 2020, **32**, 1748–1759.

7 S.-S. Wang, W.-T. Chen, Y. Li, J. Wang, H.-S. Sheu and R.-S. Liu, Neighboring-Cation Substitution Tuning of Photoluminescence by Remote-Controlled Activator in Phosphor Lattice, *J. Am. Chem. Soc.*, 2013, **135**, 12504–12507.

8 G. J. Hoerder, M. Seibald, D. Baumann, T. Schröder, S. Peschke, P. C. Schmid, T. Tyborski, P. Pust, I. Stoll, M. Bergler, C. Patzig, S. Reißaus, M. Krause, L. Berthold, T. Höche, D. Johrendt and H. Huppertz, $Sr[Li_2Al_2O_2N_2]$:Eu^{2+} - A High Performance Red Phosphor to Brighten the Future, *Nat. Commun.*, 2019, **10**, 1824.

9 P. Pust, V. Weiler, C. Hecht, A. Tücks, A. S. Wochnik, A.-K. Henß, D. Wiechert, C. Scheu, P. J. Schmidt and W. Schnick, Narrow-Band Red-Emitting $Sr[LiAl_3N_4]$:Eu^{2+} as a Next-Generation LED Phosphor Material, *Nat. Mater.*, 2014, **13**, 891–896.

10 T. Senden, R. van Dijk-Moes and A. Meijerink, Quenching of the Red Mn^{4+} Luminescence in Mn^{4+}-Doped Fluoride LED Phosphors, *Light: Sci. Appl.*, 2018, **7**, 8.

11 W. B. Park, S. P. Singh, M. Kim and K.-S. Sohn, Phosphor informatics based on confirmatory factor analysis, *ACS Comb. Sci.*, 2015, **17**, 317–325.

12 S. Li, Y. Xia, M. Amachraa, N. T. Hung, Z. Wang, S. P. Ong and R.-J. Xie, Data-Driven Discovery of Full-Visible-Spectrum Phosphor, *Chem. Mater.*, 2019, **31**, 6286–6294.

13 Z. Wang, I.-H. Chu, F. Zhou and S. P. Ong, Electronic Structure Descriptor for the Discovery of Narrow-Band Red-Emitting Phosphors, *Chem. Mater.*, 2016, **28**, 4024–4031.

14 M. Amachraa, Z. Wang, C. Chen, S. Hariyani, H. Tang, J. Brgoch and S. P. Ong, Predicting thermal quenching in inorganic phosphors, *Chem. Mater.*, 2020, **32**, 6256–6265.

15 Z. Wang, J. Ha, Y. H. Kim, W. B. Im, J. McKittrick and S. P. Ong, Mining Unexplored Chemistries for Phosphors for High-Color-Quality White-Light-Emitting Diodes, *Joule*, 2018, **2**, 914–926.

16 Y. Zhuo, A. M. Tehrani, A. O. Oliynyk, A. C. Duke and J. Brgoch, Identifying an efficient, thermally robust inorganic phosphor host via machine learning, *Nat. Commun.*, 2018, **9**, 4377.

17 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, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**, 011002.

18 Y. Zhuo, S. Hariyani, S. You, P. Dorenbos and J. Brgoch, Machine learning 5d-level centroid shift of Ce^{3+} inorganic phosphors, *Appl. Phys.*, 2020, **128**, 013104.

19 M. G. Brik, V. Jarý, L. Havlák, J. Bárta and M. Nikl, Ternary sulfides $ALnS_2$:Eu^{2+} (A = Alkaline Metal, Ln = rare-earth element) for lighting: Correlation between the host structure and Eu^{2+} emission maxima, *Chem. Eng. J.*, 2021, **418**, 129380.

20 F. Yang, Y. Wang, X. Jiang, B. Lin and R. Lv, Optimized Multimetal Sensitized Phosphor for Enhanced Red Up-Conversion Luminescence by Machine Learning, *ACS Comb. Sci.*, 2020, **22**, 285–296.

21 G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, Data Mined Ionic Substitutions for the Discovery of New Compounds, *Inorg. Chem.*, 2011, **50**, 656–663.
22 G. Hautier, A. Jain, S. P. Ong, B. Kang, C. Moore, R. Doe and G. Ceder, Phosphates as Lithium-Ion Battery Cathodes: An Evaluation Based on High-Throughput ab Initio Calculations, *Chem. Mater.*, 2011, **23**, 3495–3508.
23 G. Hautier, A. Jain, T. Mueller, C. Moore, S. P. Ong and G. Ceder, Designing Multi electron Lithium-Ion Phosphate Cathodes by Mixing Transition Metals, *Chem. Mater.*, 2013, **25**, 2064–2074.
24 L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson and L. A. Curtiss, Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening, *J. Phys. Chem. Lett.*, 2015, **6**, 283–291.
25 J. West, D. Ventura and S. Warnick, *Spring Research Presentation: A Theoretical Foundation for Inductive Transfer*, College of Physical and Mathematical Sciences, 2007.
26 D. Jha, K. Choudhary, F. Tavazza, W.-K. Liao, A. Choudhary, C. Campbell and A. Agrawal, Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning, *Nat. Commun.*, 2019, **10**, 5316.
27 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, *npj Comput. Mater.*, 2015, **1**, 15010.
28 G. Pilaniaa, J. E. Gubernatis and T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, *Comput. Mater. Sci.*, 2017, **129**, 156–163.
29 J. W. Lee, W. B. Park, B. D. Lee, S. Kim, N. H. Goo and K.-S. Sohn, Dirty engineering data-driven inverse prediction machine learning model, *Sci. Rep.*, 2020, **10**, 1–14.
30 M. Pinsky and D. Avnir, Continuous Symmetry Measures. 5. The Classical Polyhedra, *Inorg Chem.*, 1998, **37**, 5575–5582.
31 O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, *Nat. Commun.*, 2017, **8**, 15679.
32 T. Xie and J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, *Phys. Rev. Lett.*, 2018, **120**, 145301.
33 R. E. A. Goodall and A. A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry, *Nat. Commun.*, 2020, **11**, 6280.
34 S. Takemura, T. Takeda, T. Nakanishi, Y. Koyama, H. Ikeno and N. Hirosaki, Dissimilarity measure of local structure in inorganic crystals using Wasserstein distance to search for novel phosphors, *Sci. Technol. Adv. Mater.*, 2021, **22**, 185–193.
35 D. W. Marquardt, Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation, *Technometrics*, 1970, **12**, 591–612.
36 R. Tibshirani, Regression Shrinkage and Selection via the Lasso, *J. R. Statist. Soc. B*, 1996, **58**, 267–288.
37 H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B*, 2005, **67**, 301–320.
38 E. A. Nadaraya, On Estimating Regression, *Theory Probab. Its Appl.*, 2006, **9**, 141–142.
39 B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least Angle Regression, *Ann. Statist.*, 2004, **32**, 407–499.
40 G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Wiley, New York, USA, 1992.
41 D. P. Wipf and S. S. A. Nagarajan, New view of automatic relevance determination, *Adv. Neural Inf. Process. Syst.*, 2008, **20**, 1625–1632.
42 T. K. Ho, The random subspace method for constructing decision forests, *IEEE Trans, Pattern Anal. Mach. Intell.*, 1998, **20**, 832–844.
43 Y. Freund and R. E. Schapire, A short introduction to boosting, *Trans. Jpn. Soc. Artif. Intell.*, 1999, **14**, 771–780.
44 J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Statist.*, 2001, **29**, 1189–1232.
45 T. Chen and C. Guestrin, In XGBoost: In A Scalable Tree Boosting System, *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, Aug, 2016.
46 N. S. Altman, An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, *Am. Stat.*, 1992, **46**, 175–185.
47 C. Cortes and V. Vapnik, Support-vector networks, *Mach. Learn.*, 1995, **20**, 273–297.
48 G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, USA, 1990.
49 A. Höskuldsson, PLS regression methods, *J. Chemom.*, 1988, **2**, 211–228.
50 F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, *Psychol. Rev.*, 1958, **65**, 386–408.
51 J. Xiong, S.-Q. Shi and T.-Y. Zhang, A machine-learning approach to predicting and understanding the properties of amorphous metallic alloys, *Mater. Des.*, 2020, **187**, 108378.
52 E. Frank, M. A. Hall and I. H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, USA, 2016.
53 Y. Wang, Y. Tian, T. Kirk, O. Laris, J. H. Ross, R. D. Noebe, V. Keylin and R. Arróyave, Accelerated design of Fe-based soft magnetic materials using machine learning and stochastic optimization, *Acta Mater.*, 2020, **194**, 144–155.
54 C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman and Y. Su, Machine learning assisted design of high entropy alloys with desired property, *Acta Mater.*, 2019, **170**, 109–117.
55 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, In Taking the human out of the loop: a review of Bayesian optimization, *Proc. IEEE*, 2016, **104**, 148–175.
56 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss

and V. J. Dubourg, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
57 D. M. Allen, The Relationship between Variable Selection and Data Agumentation and a Method for Prediction, *Technometrics*, 1974, **16**, 125–127.
58 M. Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, *J. R. Stat. Soc.*, 1974, **36**, 111–147.
59 M. Stone, An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, *J. R. Stat. Soc.*, 1977, **39**, 44–47.
60 J. P. Perdew and M. Levy, Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities, *Phys. Rev. Lett.*, 1983, **51**, 1884–1887.
61 L. J. Sham and M. Schlüter, Density-Functional Theory of the Energy Gap, *Phys. Rev. Lett.*, 1983, **51**, 1888–1891.
62 J. M. Crowley, J. Tahir-Kheli and W. A. Goddard, Resolution of the Band Gap Prediction Problem for Materials Design, *J. Phys. Chem. Lett.*, 2016, **7**, 1198–1203.
63 H. Pan, A. M. Ganose, M. Horton, M. Aykol, K. A. Persson, N. E. R. Zimmermann and A. Jain, Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures, *Inorg. Chem.*, 2021, **60**, 1590–1603.
64 L. Pauling, *The Nature of the Chemical Bond and Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*, Cornell University, Ithaca, NY, 1960, pp. 65–105.
65 R. D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chaleogenides, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 1976, **32**, 751–767.
66 J. Henseler, C. Ringle and R. Sinkovics, The use of partial least squares path modeling in international marketing, *Adv. Int. Mark.*, 2009, **20**, 277–320.
67 I. Kononenko, Bayesian neural networks, *Biol. Cybern.*, 1989, **61**, 361–370.
68 The NOMAD Laboratory, A European Centre for Excellence, https://nomad-coe.eu/, (accessed March 2017).
69 S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo and O. Levy, Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations, *Comput. Mater. Sci.*, 2012, **58**, 227–235.
70 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), *JOM*, 2013, **65**, 1501–1509.
71 A. Lehman, N. O'Rourke, L. Hatcher and E. J. Stepanski, *Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide*, SAS Institute, North Carolina, 2005.
72 P. Dorenbos, 5d-level energies of Ce^{3+} and the crystalline environment. I. Fluoride compounds, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2000, **62**, 15640–15649.
73 P. Dorenbos, 5d-level energies of Ce^{3+} and the crystalline environment. II. Chloride, bromide, and iodide compounds, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2000, **62**, 15650–15659.
74 P. Dorenbos, 5d-level energies of Ce^{3+} and the crystalline environment. III. Oxides containing ionic complexes, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2001, **64**, 125117.
75 P. Dorenbos, Relation between Eu^{2+} and Ce^{3+} f $\leftrightarrow$ d-transition energies in inorganic compounds, *J. Phys.: Condens. Matter*, 2003, **15**, 4797–4807.
76 C. A. Morrison, Host dependence of the rare-earth ion energy separation $4f^N$–$4f^{N-1}$ nl, *J. Phys. Chem.*, 1980, **72**, 1001–1002.
77 B. F. Aull and H. P. Jenssen, Impact of ion-host interactions on the 5d-to-4f spectra of lanthanide rare-earth-metal ions. I. A phenomenological crystal-field model, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1986, **34**, 6640–6646.
78 J. D. Axe and G. Burns, Influence of Covalency upon Rare-Earth Ligand Field Splittings, *Phys. Rev.*, 1966, **152**, 331–340.
79 S. Shionoya, W. M. Yen and H. Yamamoto, *Phosphor Handbook*, CRC Press, Bpca Raton, Florida, 2018.