Electronic Supplementary Material (ESI) for Materials Advances.

Supplementary Information

Rapid Discovery of New Eu^{2+}-Activated Phosphors with a Designed Luminescence Color by a Data-Driven Approach

Yukinori Koyama,*[a] Hidekazu Ikeno,[b] Masamichi Harada,[c] Shiro Funahashi,[c] Takashi Takeda[c] and Naoto Hirosaki[c]

[a] Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan

[b] Department of Materials Science, Graduate School of Engineering, Osaka Metropolitan University, Sakai, Osaka 599-8570, Japan

[c] Research Center for Functional Materials, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan

* Email: KOYAMA.Yukinori@nims.go.jp

Table S1. List of elemental features used for the general-purpose features. The elemental features were obtained from the XenonPy package (Ref. 23).

Elemental feature
Atomic number
Period
Group
Atomic mass
Number of valence electrons
Number of valence s electrons
Number of valence p electrons
Number of valence d electrons
Number of valence f electrons
Number of unoccupied valence states
Number of unoccupied valence s states
Number of unoccupied valence p sates
Number of unoccupied valence d states
Number of unoccupied valence f states
Atomic radius
Covalent radius
Van de Waals radius
Electronegativity
Electron affinity
First ionization energy
Mendeleev number
Polarizability

Table S2. List of statistics used for the general-purpose features. f_i and w_i ($\sum_i w_i = 1$) denote an elemental feature and atomic fraction of element i, respectively. n denotes the number of elements.

Statistic	**Equation**
Weighted arithmetic mean	$f_{\text{mean}} = \sum_{i=1}^{n} w_i f_i$
Weighted geometric mean	$f_{\text{g-mean}} = \prod_{i=1}^{n} f_i^{w_i}$
Weighted harmonic mean	$f_{\text{h-mean}} = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{f_i}}$
Weighted standard deviation	$f_{\text{sd}} = \sqrt{\sum_{i=1}^{n} w_i (f_i - f_{\text{mean}})^2}$
Minimum	$f_{\text{min}} = \min\{f_i\}$
Maximum	$f_{\text{max}} = \max\{f_i\}$
Range	$f_{\text{range}} = f_{\text{max}} - f_{\text{min}}$

Table S3. Feature-selection and regression pipeline, parameter ranges and optimized values of the a) ridge, b) automatic relevance determination (ARD), c) random forest (RF), d) gradient boosted regression trees (GB), and e) bagging of GB models. Classes and functions in the scikit-learn package are listed without their module names.

a) Ridge

Estimator	**Parameter**	**Range**	**Optimized value**
VarianceThreshold	threshold	Fixed	1.0e-7
StandardScaler			
SelectKBest	score_func	Fixed	mutual_info_regression
	k	100, 150, …, 350	300
RFE	estimator	Fixed	Ridge (default parameters)
	n_features_to_select	10, 20, …, 100	90
	step	Fixed	10
Ridge	alpha	[1.0e-6, 1.0e+6] (log-uniform)	35.1

b) ARD

Estimator	Parameter	Range	Optimized value
VarianceThreshold	threshold	Fixed	1.0e-7
StandardScaler			
SelectKBest	score_func	Fixed	mutual_info_regression
	k	100, 150, …, 350	350
RFE	estimator	Fixed	ARDRegression (default parameters)
	n_features_to_select	10, 20, …, 100	60
	step	Fixed	10
ARDRegression	alpha_1	[0.0, 1.0] (uniform)	0.734
	alpha_2	[1.0e-6, 1.0e+6] (log-uniform)	1.73e-5
	lambda_1	[0.0, 1.0] (uniform)	0.335
	lambda_2	[1.0e-6, 1.0e+6] (log-uniform)	0.494
	threshold_lambda	[1.0e+2, 1.0e+6] (log-uniform)	4.65e+4

c) RF

Estimator	Parameter	Range	Optimized value
VarianceThreshold	threshold	Fixed	1.0e-7
StandardScaler			
SelectKBest	score_func	Fixed	mutual_info_regression
	k	100, 150, …, 350	350
RFE	estimator	Fixed	RandomForestRegressor (default parameters)
	n_features_to_select	10, 20, …, 100	100
	step	Fixed	10
RandomForestRegressor	max_depth	1, 2, …, 20	13
	min_samples_leaf	1, 2, 3	1
	n_estimators	50, 60, …, 200	180

d) GB

Estimator	Parameter	Range	Optimized value
VarianceThreshold	threshold	Fixed	1.0e-7
StandardScaler			
SelectKBest	score_func	Fixed	mutual_info_regression
	k	100, 150, …, 350	350
RFE	estimator	Fixed	GradientBoostingRegressor (default parameters)
	n_features_to_select	10, 20, …, 100	70
	step	Fixed	10
GradientBoostingRegressor	learning_rate	[0.01, 0.5] (log-uniform)	0.111
	max_depth	1, 2, …, 5	3
	n_estimators	100, 200, …, 1000	900

e) Bagging of GB

Estimator	Parameter	Range	Optimized value
VarianceThreshold	threshold	Fixed	1.0e-7
StandardScaler			
SelectKBest	score_func	Fixed	mutual_info_regression
	k	100, 150, …, 350	350
RFE	estimator	Fixed	GradientBoostingRegressor (default parameters)
	n_features_to_select	10, 20, …, 100	100
	step	Fixed	10
BaggingRegressor	base_estimator	Fixed	GradientBoostingRegressor
	n_estimators	Fixed	25
GradientBoostingRegressor (base_estimator)	learning_rate	[0.01, 0.5] (log-uniform)	0.213
	max_depth	1, 2, …, 5	3
	n_estimators	100, 200, …, 1000	400

Table S4. Mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) of the machine learning on emission peak energy using the bagging of the gradient boosted regression trees method for the training and validation data in the cross validation. The scores were averaged among the folds of the cross validation. Standard deviations among the folds are shown in parentheses.

Score	**Training**	**Validation**
MAE / eV	0.05 (0.00)	0.13 (0.03)
RMSE / eV	0.07 (0.00)	0.16 (0.05)
R^2	0.97 (0.00)	0.78 (0.16)

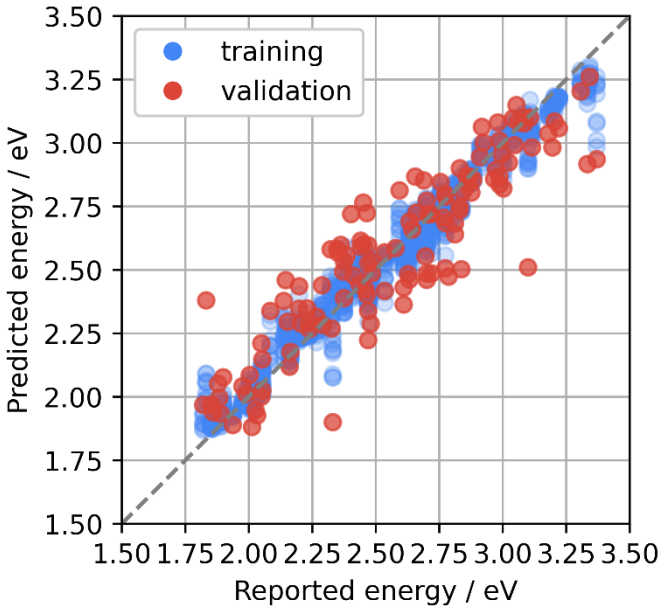
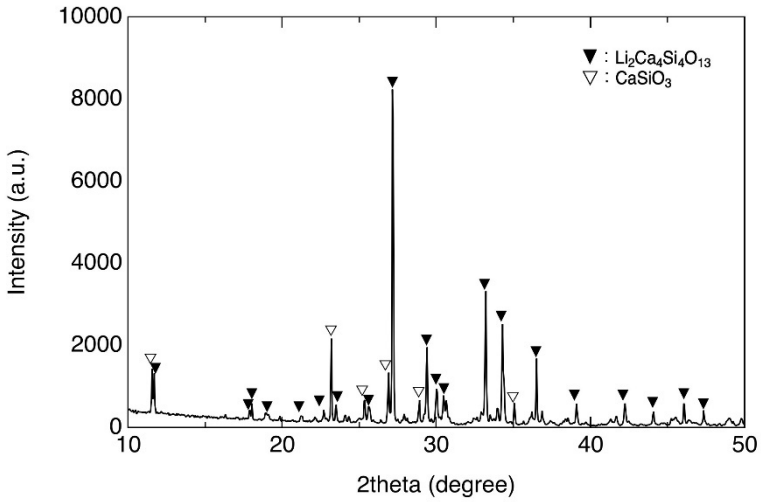
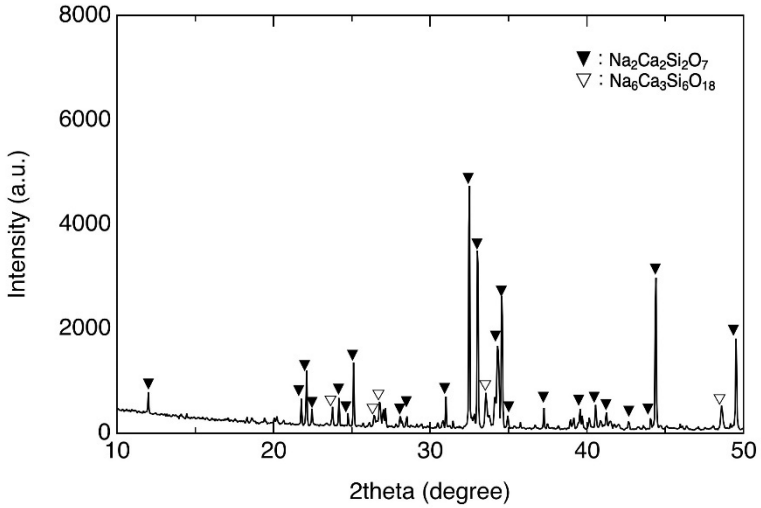


Figure S1. Predicted emission peak energies with respect to reported values for the training (blue) and validation (red) data in the cross validation using the bagging of the gradient boosted regression trees method.

a) $Li_2Ca_4Si_4O_{13}$



b) $Na_2Ca_2Si_2O_7$



c) $SrLaGaO_4$

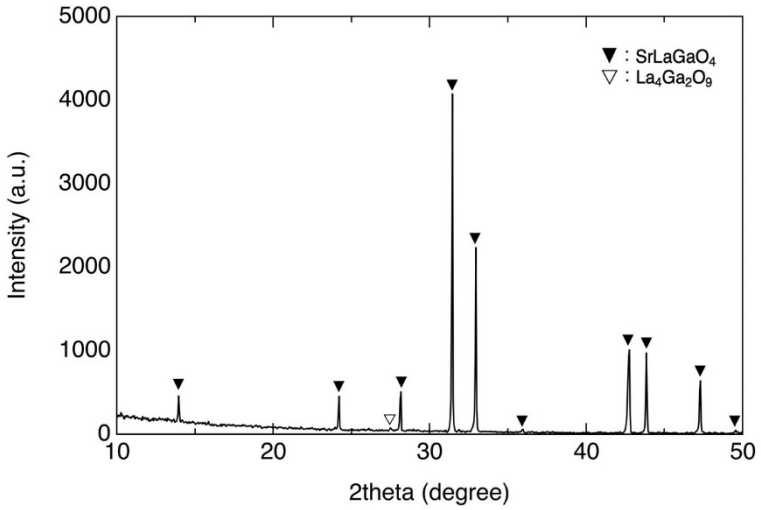


Figure S2. XRD patterns of powder samples of Eu-doped (a) $Li_2Ca_4Si_4O_{13}$, (b) $Na_2Ca_2Si_2O_7$, and (c) $SrLaGaO_4$.