scientific reports

OPEN

Predicting refractive index of inorganic compounds using machine learning

Elham Einabadi[1] & Mahdi Mashkoori[1,2]✉

Refractive index (RI) is one of the most important optical properties of materials. Due to the high importance of this physical parameter, there has always been a demand to find a method that provides the most optimal estimation. In this research, we utilize experimentally measured *RI* values of 272 inorganic compounds to build a machine learning model capable of predicting the *RI* of materials with low computational cost. Considering the significant relationship between the band gap and *RI*, we select this parameter as a predictor. In addition to the band gap, the atomic properties related to the building elements of the compounds form our data set in this work. To find the most optimal model and set of suitable predictors, we examine our data in four categories with 1, 5, 10, and 21 features. In addition, we compare the predicted *RIs* of 6 different independent regression methods, namely, ordinary least squares (OLSR), Gaussian process (GPR), support vector (SVR), random forest (RFR), gradient boosted trees (GBTR), and extremely randomized trees regression(ERTR). We notice that ERTR predicts *RI* with the highest accuracy compared to other regression methods. The prediction strength of our model excels in empirical relations and provides accurate results for a wide range of *RIs*. Thus, we demonstrate the high potential of machine learning methods for evaluating the *RI*, especially when it comes to providing an estimation of a desired physical quantity.

Refractive index (RI) is one of the most fundamental optical properties of materials that determines the propagtion velocity of electromagnetic waves[1]. It will be different depending on the type of symmetry of the inorganic compound and the wavelength of the light[2]. Knowing the *RI* of inorganic compounds for each specific wavelength is essential to understanding the behavior of these materials and is widely used for designing optical devices in industry[3]. For instance, knowing *RI* is vital in manufacturing optical devices like switches, filters, and modulators[4]. Therefore, many efforts have been made to find an empirical formula that expresses the *RI* in terms of other properties of materials[5,6]. Moreover, the accuracy and generalizability of the empirical formulas to obtain the *RI* have improved over time due to advances in measurement techniques and tools. However, these estimates still have their limitations. Attempts have been made to calculate *RI* since the mid-19th century, and several relations have been suggested. One of the first attempts is the Lorentz-Lorenz formula to evaluate the *RI* for different compounds. In this equation, the *RI* is estimated only based on the material density[7,8]:

$$\frac{n^2-1}{n^2+1}=\frac{4\pi}{3}N\alpha, \tag{1}$$

where n is the refractive index, N is the number density, and α is the polarizability coefficient. This model does not include the band structure of materials and consequently leads to its failure in several cases. Considering the significant relationship between *RI* and band gap (E_g), numerous empirical formulas were introduced between E_g and *RI* such as Moss formula[9]:

$$n^4E_g = 95\ eV, \tag{2}$$

and Ravindra formula[10]:

$$n = 4.084 - 0.62\ E_g. \tag{3}$$

[1]Department of Physics, K.N. Toosi University of Technology, P. O. Box 15875-4416, Tehran, Iran. [2]School of Physics, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5531, Tehran, Iran. ✉email: mahdi.mashkoori@kntu.ac.ir

These empirical formulas have some weaknesses. For example, Eq.3 shows very large deviations for small ($E_g \leq 0.3$ ev) and large ($E_g \geq 3.5$ ev) band gaps. Moreover, this relation fails when $n \geq 4.1$[11,12]. Thus, the above-mentioned equations often yield reasonable predictions only for specific materials and in a specific range of *RI*. As a treatment for this weakness, Herve et al. presented an alternative relation between *RI* and E_g taking into account the observed deviations in small and large E_g values[11]:

$$n = \sqrt{1 + \left(\frac{13.6}{E_g + 3.4}\right)^2}. \tag{4}$$

This relation works well for many optoelectronic materials, but it shows low accuracy for materials from group IV-VI[13,14].

In another work, Reddy[12] proposed an additional empirical relation between *RI* and optical electronegativity($\Delta\chi^*$) as follows:

$$n = -\ln\{0.102\Delta\chi^*\}. \tag{5}$$

Duffy[15] earlier introduced a relation between E_g and $\Delta\chi^*$ using the following equation:

$$\Delta\chi^* = 0.2688 E_g. \tag{6}$$

Usually, a theoretically predicted *RI* is compared with experimental *RI* that is measured using interferometric, ellipsometric, spectroscopic, and methods using prisms[16]. The challenges in experimental evaluation of *RI* and the complexity of the problem require the use of new approaches to predict *RI*, assisted by growing computational resources and better algorithms. For instance, as an alternative to empirical formulas, machine learning (ML) provides a numerical approach to evaluate the physical properties of materials. As an instance, Lee et al. predicted the E_g of inorganic materials using different ML methods, namely ordinary least squares regression (OLSR), least absolute shrinkage and selection operator, and nonlinear support vector regression (SVR)[17]. Using a data set that consists of the E_g of 270 different compounds calculated using DFT $(\mathrm{G}_0 \quad \mathrm{W}_0)$ and, 18 fundamental information of constituent elements as predictors, they show that ML-assisted algorithms predict the E_g with reasonable accuracy.

Generally, engineering new compounds and discovering materials with desirable properties has always been an important task for the scientific community[18–20]. Considering the number of elements in the periodic table, a huge number of compounds can be identified with their unique physical features and each of these properties is important for specific application purposes. An accurate and accessible machine learning (ML) model can greatly accelerate the identification of new functional materials[21]. Because providing an estimation for features of interest for a given structure using ML is numerically much less expensive compared to ab initio methods. Furthermore, ML has been successful in predicting the physical properties of materials and, we observe rapid advancements in data science and material informatics. Therefore, exploring machine learning methods to predict *RI* represents a promising step forward in this field. Moreover, providing an accurate estimation of the optical properties in general and specially *RI* of materials is very important for various applications including, but not limited to, laser optics, optical glasses, and optical fibers.

Recently, some efforts have been made in the prediction of *RI* using ML[22–24]. For example, Kang et al. used extreme machine learning (ELM) and multiple linear regression (MLR) algorithms to predict *RI*. The *RI* investigated in this research was in the range of 1.36 to 1.6 which covers a narrow window of *RI*s for optical materials. Focusing on organic compounds, Lightstone et al. also formed a data set including 527 unique polymers and estimated the *RI* of these compounds using the Gaussian process regression algorithm. This group examined polymers with *RI* in the range of 1.3 to 2. Aiming to cover both organic and inorganic materials, Zhao et al. developed a model using three machine learning models, Support Vector Regression (SVR), Random Forest Regression (RFR), and Gaussian Process Regression (GPR). In this work, a data set of materials with 49,076 experimental *RI* values for 6,721 compounds has been studied. The results of these investigations were more accurate compared to other empirical formulas like Moss and Ravindra.

Despite this pioneering research, there is still much room for improving the results, and this goal can be achieved by choosing more suitable machine learning techniques and better and more accessible predictors. In this paper, we examine *RI* prediction models using OLSR, GPR, SVR, RFR, GBTR, and ERTR methods. Moreover, we prepare our data set aiming at inorganic materials from a broad range of *RI*s. Specifically, our data covers *RI* in the range of 1.3 to 4. By selecting physically relevant predictors, namely the band gap and elemental properties of constituent elements, and in spite of the fact that our data set includes only 272 inorganic compounds, we achieve a notably low error in predicting *RI*s and our model outperforms the empirical formulas.

Methodology

Machine learning models

Machine learning (ML) is a subfield of artificial intelligence technology and one of the most attractive and dynamic fields of modern research and application. ML methods are based on enabling computers to identify and infer patterns in data without being explicitly programmed and to build models with the ability to make predictions that do not explicitly follow predefined rules and models[25–28]. Regression is a supervised learning method that obtains a relation between a target attribute and predictors. For a given training data set, it is necessary to select appropriate predictors and optimal ML methods. In this work, we select six ML methods to predict the *RI* of 272 different compounds. These methods are chosen from a wide range of ML methods based on their performance in the previous studies[17,24,29]:

As a simple, yet important ML method, OLSR is a method by which the norm of the cost function reaches its minimum value[30]:

$$L = \min_{w} \|Xw - y\|_2^2 = \min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2. \tag{7}$$

In this relation, *w* represents the vector of weights, x_i is the observed input vector, and y_i is the observed target value. In this method, the predictive function $f(x_i, w) = w^T x_i$ creates a hyperplane in the feature space. Therefore, minimizing the error of this method means minimizing the sum of squares of the difference between the actual value of y_i and the predicted value, for all the points of x_i relative to this hyperplane.

Moreover, compared to OLSR, the Support Vector Machine (SVM) is a more advanced and powerful ML method. SVM is based on the theory of statistical and dynamical learning. The main idea of SVM was first presented in the 1960s by Vepnik et al.[31]. In the following years, this model was expanded to regression problems under the name Support Vector Regression (SVR)[32]. This method strikes a balance between model complexity and prediction error and performs well in high-dimensional data analysis. The function of SVR is based on creating a hyperplane with support vectors where optimization is done according to the support vectors and is independent of the input data dimensionality[33]. Additionally, the parameters controlling the regression function are obtained by solving the following optimization problem[34]:

$$Minimize : \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{8}$$

$$\begin{aligned} Subject\,to : &\{y_i - x_i w - b \le \varepsilon + \xi_i\}, \\ &\{x_i w + b - y_i \le \varepsilon + \xi_i^*\}. \end{aligned}$$

where ξ_i and ξ_i^* are slack variables $(\xi_i, \xi_i^* \ge 0)$, C is regularization constant and the weight vector denoted by *w*. The optimization problem can be solved by constructing a Lagrange function from the primal objective function and solving the so-called dual optimization problem. Moreover, the parameters of the optimal function for the nonlinear regression of the support vector can be rewritten with the following relations:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b, \tag{9}$$

$$b = y_i + \varepsilon - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_j, x_i), \tag{10}$$

where $k(x_j, x_i)$ represent the kernel function, and α_i and α_i^* represent the Lagrangian coefficients. The use of kernels is one of the most common approaches in the support vector method, because in this case, linear optimization is performed in the kernel space, and there is no need to create a high-order separating hyperplane in the input data space, which is a very complicated method. The kernel we use in this method is the Radial Basis Function (RBF) kernel:

$$k(x, x') = \sigma_f^2 exp\left(-\frac{1}{2l^2}\|x - x'\|^2\right), \tag{11}$$

where, signal variance σ_f^2 and length scale *l* are two hyperparameters of this kernel[35].

Considering the next ML method that we use in this paper, GPR is a non-parametric Bayesian method[35–38]. Gaussian process (GP) is a random process in the form of a set $\mathcal{F}$ of random variables $F_{x_1}, F_{x_2}, \ldots$. Each subset of these variables has a common multivariate Gaussian distribution[37]. In fact, GP is a generalization of the Gaussian probability distribution. Kernel is an important element in GP that determines its posterior and prior shape. By assuming a kernel, also known as the covariance function, similar data points should lead to similar target values. Choosing an appropriate kernel is based on assumptions such as the kernel's consistency with expected patterns in the data. In this method, a GP is assumed, which can be shown using a mean function *m*(*x*) and covariance function $k(x, x')$:

$$f(x) \sim GP(m(x), k(x, x')), \tag{12}$$

where mean and covariance are defined as follows:

$$m(x) = E[f(x)] \tag{13}$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))] \tag{14}$$

The kernel we employ in this work is the RBF kernel, which is one of the most common kernels used in GPR. Ensemble methods are advanced machine learning techniques that enhance accuracy and robustness by integrating the predictions of multiple models, rather than relying on a single model. Notable examples of ensemble methods include Random Forest (RF)[39], Gradient Boosted Trees (GBT)[40,41], and Extremely Randomized Trees (ERT)[42]. Among these, RF is one of the most widely used and adaptable ML algorithms that is used for both regression and classification problems. Decision tree model, also called classification and regression tree (CART), is a non-parametric model that was first introduced by Breiman et al. in 1984[43]. CART is of interest due to its simplicity and low computational cost, as well as, interpretability and the possibility of graphical representation[44]. On the other hand, a decision tree is a model that possesses low predictive accuracy due to its inherent high variance. However, several CART models can be combined to create an ideal ensemble model, and hence it is called a "random forest". More precisely, RF consists of several independent and uncorrelated decision trees, each tree alone implements a classification or regression algorithm according to the type of problem. In the classification method, the output of RF is the class that obtains the highest count of votes from the set of decision trees. As for the regression method (RFR), the final output is the mean of outputs of all decision trees[45–47].

Database

In the present study, a data set of 272 inorganic compounds is created by extracting data from independent scientific literatures[2,12,48–66]. This data set includes experimental *RI* and E_g values of inorganic compounds, along with additional information regarding their constituent elements. The range of *RI* values is from 1.3 to 4, with most falling between 1.5 and 2.5. The distribution of these *RI*s is illustrated in Fig. 1a.

Besides, these compounds with their *RI* and E_g values, can be found in the supplementary materials. Moreover, the characteristics of the constituent elements of the compounds used in the data set include period *p* within the periodic table, atomic number *Z*, atomic mass *m*, electronegativity χ, the first ionization energy *I*, atomic radius *r*, melting point T_M, boiling point T_B, density *D*, and conductivity σ. For a given compound, element-specific predictors are calculated in the form of mean, $\langle c \rangle$, and standard deviation, σ_c, as follows:

$$\langle C \rangle = \sum_{k=1}^{N} x_k C_k, \tag{15}$$

$$\sigma_C = \sqrt{\sum_{k=1}^{N} x_k \left(C_k - \langle C \rangle\right)^2}, \tag{16}$$

where C_k and x_k are the values of the basic variables of each constituent element and the contribution of each of them in the compound, respectively. Also, *N* represents the number of elements in the compound. So, all these derived parameters form a set of 21 predictors which we use in the following sections to predict *RI*.

Results and discussion

In this section, we present our results for the prediction of *RI* using different regression methods as described in above. We show that using simple regression models can be useful, yet leads to low accuracy in predicting the

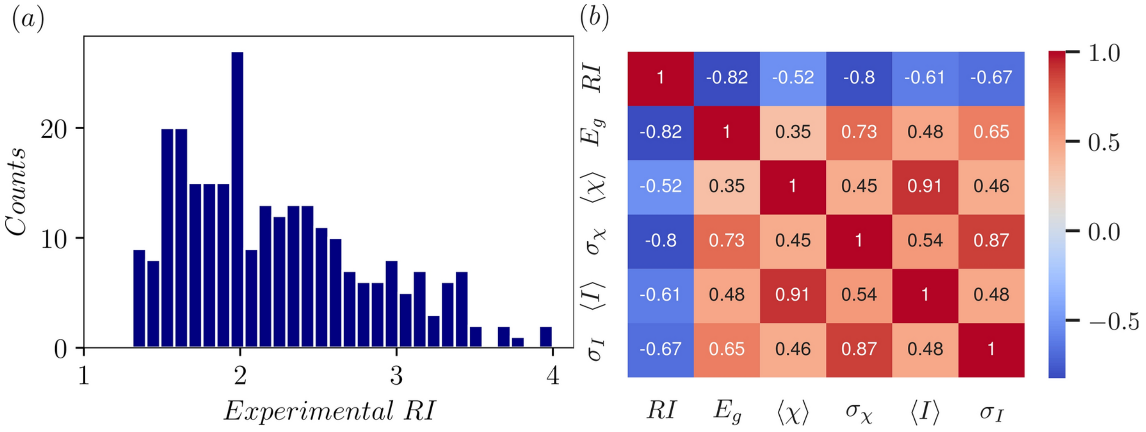

Fig. 1. (**a**) Distribution of refractive indices(RIs) in the data set, including 272 inorganic compound. (**b**) Correlations between predictors and refractive index(RI). E_g, $\langle\chi\rangle$, σ_χ, $\langle I\rangle$, and σ_I represent mean and standard deviation of electronegativity and the first ionization energy, respectively. The negativity of correlation indicates an inverse relation between the variables.

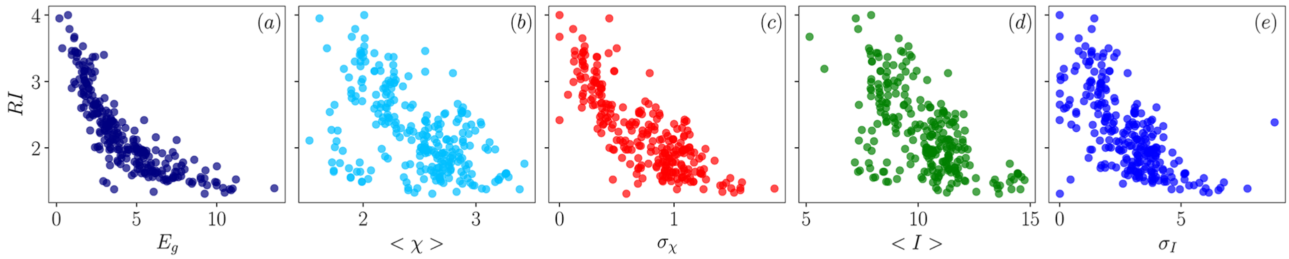

Fig. 2. Relation between predictors and refractive index (*RI*). (**a**) Illustrates the relation between *RI* and E_g, while (**b**,**c**) show *RI* as a function of the mean and standard deviation of electronegativity (χ), respectively. (**d**,**e**) illustrate the dependence of *RI* on the mean and standard deviation of first ionization energy (*I*).

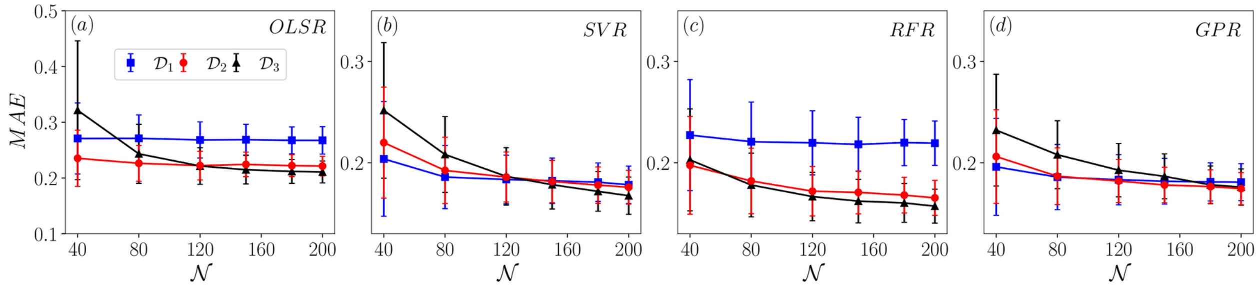

Fig. 3. Dependence of MAE on the number of training data ($\mathcal{N}$) using OLSR (**a**), SVR (**b**), RFR (**c**), and GPR (**d**) methods for three groups of data sets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$. Error bars represent the standard deviation for 500 different trials.

target parameter. However, by using more advanced ML methods, we demonstrate a more accurate prediction of the *RI* for the compounds in our data set.

As the first step, we assess the relation between 21 predictors and the target parameter, namely *RI*. Considering the correlation between predictors and *RI*, the most correlated parameters to *RI* appear to be the E_g, the electronegativity (χ), and the first ionization energy (*I*) as depicted in the heatmap of Fig. 1b. Considering the first row, this figure shows that *RI* has a notable correlation (-0.82) with E_g and σ_χ. The negativity of correlation hints that *RI* should be related to the inverse of E_g and other predictors. Furthermore, we notice a higher correlation between *RI* and the standard deviation of predictors, namely σ_χ and σ_I, in comparison to their mean, $\langle\chi\rangle$ and $\langle I\rangle$. Also, we note high correlations among the predictors as well. For instance the correlation between $\langle\chi\rangle$ and $\langle I\rangle$ is 0.91. Concentrating more closely on the relation between *RI* and predictors, Fig. 2 depicts the relation between *RI* and E_g, *I*, and χ for all the compounds in our data set. In agreement with the heatmap of Fig. 1b, we see *RI* inversely depends on E_g, σ_χ, and σ_I. In addition, *RI* shows a weak correlation with $\langle\chi\rangle$ and $\langle I\rangle$.

The results, demonstrate a significant association between *RI* and E_g. Moreover, there is a notable correlation between *RI* and both the electronegativity and first ionization energy. This collides with our physical intuition about importance of electronic excitation of materials in studying optical properties. Therefore, we generated 3 groups of data sets; $\mathcal{D}_1$: Data set with only E_g as predictor. $\mathcal{D}_2$: Data set with 5 features including the E_g and the mean and standard deviation of the first ionization energy and electronegativity. $\mathcal{D}_3$: Data set including all 21 features. In what follows, we compare the performance of ML models using these three different data sets. Clearly, we compare the accuracy of different models for each data set as well.

In this study, the performance of ML methods is assessed through evaluation of the mean absolute error (MAE), that is:

$$MAE(y, y') = \frac{1}{n_{\text{ test}}} \sum_{i=1}^{n_{\text{ test}}} |y - y'|, \tag{17}$$

where *y* is the observed target feature, y' is the predicted target feature, and $n_{\text{ test}}$ is the number of test data.

The analysis of the MAE for ML models is illustrated in Fig. 3, where the vertical axis shows the MAE and the horizontal axis illustrates the size of the training data set. In Fig. 3a–d we implemented OLSR, SVR, RFR, and GPR models, respectively. For each model, we split the training ($\mathcal{N}$) and test ($\mathcal{N}'$) sets randomly from 272 compounds in each trial while keeping the ratio of train to test as three to one. Furthermore, we apply each model to above-mentioned data sets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ which include 1, 5, and 21 features, respectively. Each point in Fig. 3 represents an average of over 500 randomly chosen configurations of train-test splitting, and error bars illustrate the corresponding standard deviation. Clearly, this figure suggests that all ML models used in this work converge for a large enough training data set.

As can be seen in Fig. 3a, by using OLSR the error saturates very quickly for $\mathcal{D}_1$. By considering data set $\mathcal{D}_2$ with 5 predictors, we find almost the same trend with a lower saturated MAE. However, by using data set $\mathcal{D}_3$ which increases the predictors to 21, we see a slow reducing MAE trend a function of $\mathcal{N}$ that gives rise to a slightly lower MAE in comparison to $\mathcal{D}_2$. Furthermore, the standard deviation falls rapidly by increasing the size of $\mathcal{N}$. It is apparent that the OLSR method yields a minimum error of approximately 0.22 for $\mathcal{D}_2$ and $\mathcal{D}_3$. We also confirm that using principle component analysis leads to similar results. For the SVR model, we implement radial basis function (RBF) as the kernel and the results are depicted in Fig. 3b. Here, we observe MAE declines by increasing $\mathcal{N}$ and this trend hints that MAE gets saturated for larger data sets. As shown in this figure, SVR for $\mathcal{D}_1$ and $\mathcal{D}_2$ leads to an MAE ≈ 0.17 for $\mathcal{N} = 200$. Although using $\mathcal{D}_3$ gives larger MAE for small $\mathcal{N}$, by increasing $\mathcal{N}$ this data set surpasses $\mathcal{D}_1$ and $\mathcal{D}_2$ in giving lower MAE by having a larger slope. Exploiting the RFR method, we utilize 35 trees to predict *RI* and the outcome is presented in Fig. 3c. We employ the GridSearchCV technique from SciKit-learn to determine the optimal number of trees. Apparently, using only the band gap as a predictor does not give any satisfactory result. However, using $\mathcal{D}_2$ we notice a declining trend for MAE as a function of $\mathcal{N}$. Furthermore, with all 21 features selected, MAE leads to an even more accurate prediction of *RI*. Obviously, the declining slope is not saturated and this promises better performance for larger data sets. In the GPR method, akin to SVR we use RBF for the kernel and show the result in Fig. 3d. Similar to RFR, the result for $\mathcal{D}_1$ is not promising. On the other hand, considering $\mathcal{D}_2$, MAE decreases as $\mathcal{N}$ increases, and for $\mathcal{D}_3$, its declining slope is larger than the other two sets. However, for $\mathcal{N} = 200$, using $\mathcal{D}_2$ still gives more accurate prediction compared to $\mathcal{D}_3$. Given that the data set used in this work is limited in size, we expect $\mathcal{D}_3$ to outperform $\mathcal{D}_2$ for larger data sets.

Obviously, MAE for the OLSR method saturates rapidly and does not lead to accurate prediction of *RI* compared to SVR, GPR, and RFR. Consequently, as illustrated in Fig. 4a, we compare the performance of these three methods using data set $\mathcal{D}_3$ which includes all 21 features we prepared for inorganic compounds in this work. All the methods share a declining trend for MAE by increasing $\mathcal{N}$, resulting in a more accurate prediction for larger data sets. Moreover, the RFR method gives the best performance with $MAE = 0.156$. Therefore, we find the RFR method to be the optimum ML model for predicting *RI* among the models evaluated so far in this work.

We also assess the prediction strength of our trained models by calculating MAE using 10-fold cross validation for sets of 20 unseen compounds out of 272, as depicted in Fig. 4b. Here, $\mathcal{D}_1$ indicates performing the ML method for the data set with E_g as a predictor. Similarly, $\mathcal{D}_2$ and $\mathcal{D}_3$ represent the performance of the method for data sets with 5 and 21 features, respectively. Clearly, for the SVR and GPR method choosing E_g as the only predictor gives a reasonable result, and increasing the number of predictors does not lead to a significant improvement. In addition, SVR and GPR trained models show an over-fitting issue, since the MAE increases drastically by using $\mathcal{D}_3$ in comparison to $\mathcal{D}_1$. However, utilizing the RFR model for $\mathcal{D}_3$ shows a significant improvement for unseen data. Thus, the RFR model with 21 features can be introduced as a more successful model for predicting *RI* in contrast to other models examined so far. For the sake of clarity, Table 1 provides the predicted values of these 3 models for some unseen compounds. Furthermore, besides the MAE and to monitor the quality of these predictions, the mean percentage of absolute error (MAPE) for these predictions have been calculated as[17]:

$$MAPE\,(y, y') = \frac{1}{n_{\text{ test}}} \sum_{i=1}^{n_{\text{ test}}} \left| \frac{y - y'}{y} \right| \times 100, \tag{18}$$

where *y* is the observed target feature, y' is the predicted target feature, and $n_{\text{ test}}$ is the number of test data. As indicated in Tab. 1, we find that using tree-based method, gives the highest accuracy compared to other regression methods. A detailed table featuring 20 unseen compounds can be found in the supplementary materials.

The SHAP summary plot, Fig.5a illustrates the impact and importance of various features on the RFR predictions. The $x-$axis represents SHAP values, indicating the effect of each feature on the model output,

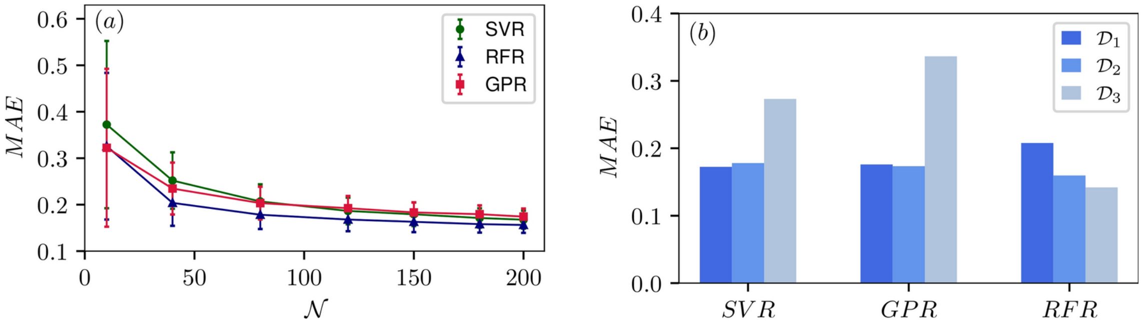

Fig. 4. (**a**) Dependence of MAE of the test set on the number of training data ($\mathcal{N}$) using SVR, RFR, and GPR models for the data set with 21 predictors ($\mathcal{D}_3$). Error bars represent the standard deviation for 500 different trials. (**b**) The prediction MAE of SVR, GPR, and RFR using 10-fold cross validation for sets of 20 unseen compounds. $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ represent data sets with 1, 5 and 21 predictors.

Compound	Experimental Value	SVR1	SVR2	SVR3	GPR1	GPR2	GPR3	RFR1	RFR2	RFR3	ERTR4	Ravindra	Moss	Herve	Reddy
AgAlS2	2.458	2.304	2.548	2.469	2.322	2.547	2.762	1.868	2.385	2.326	2.493	2.143	2.347	2.310	2.455
AgGaS2	2.390	2.473	2.675	2.511	2.488	2.677	2.680	2.349	2.456	2.392	2.469	2.391	2.428	2.433	2.592
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
ZnS	2.350	2.133	2.367	2.565	2.133	2.367	2.985	2.157	2.325	2.265	2.318	1.802	2.254	2.165	2.293
ZrO2	2.122	1.884	1.855	2.045	1.930	1.940	2.079	1.929	1.927	1.974	1.948	0.984	2.087	1.902	1.987
MAE	–	0.148	0.150	0.219	0.159	0.140	0.300	0.194	0.134	0.117	0.108	0.403	0.175	0.162	0.159
MAPE(%)	–	6.895	6.834	9.521	7.341	6.310	13.079	8.905	6.317	5.362	4.845	18.998	7.901	7.279	7.699

Table 1. The predicted refractive index values using ML models of SVR, RFR, GPR, and ERTR in addition to calculated values using empirical relations. SVR*i* represent the prediction using the SVR method with data sets $\mathcal{D}_i$. The same explanation holds for the other methods. MAE and MAPE show the errors of these methods.

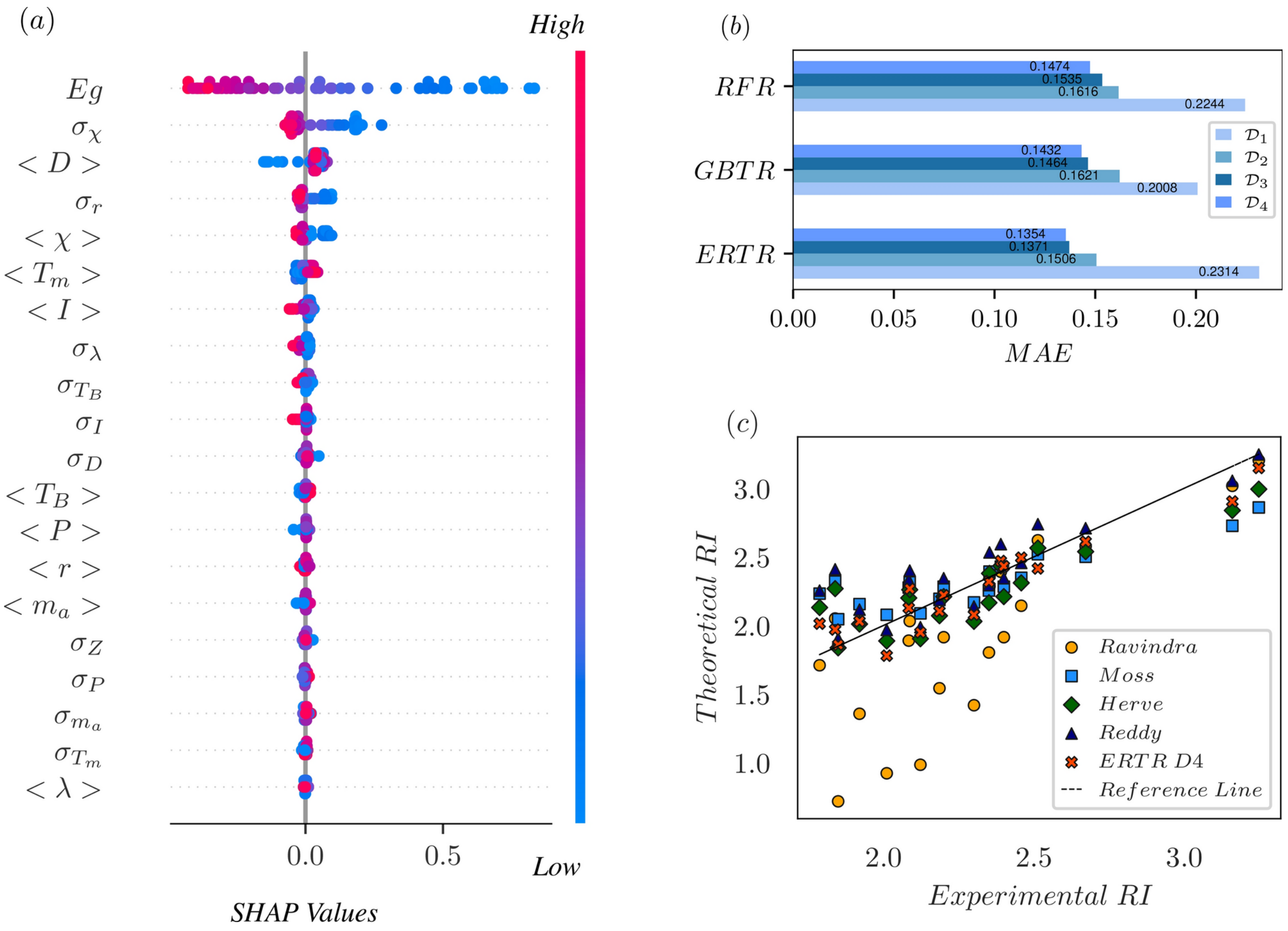

Fig. 5. (**a**) The top 20 features that impact the prediction of the RFR, arranged from highest to lowest importance. The horizontal axis shows SHAP values, indicating each feature impact on predictions. Features are ranked by importance on the vertical axis. Dots represent data instances, colored from blue (low) to red (high). (**b**) MAE of RFR, GBTR and ERTR methods for 10-fold cross-validation. $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$ represent data sets with 1, 5, 21, and 10 predictors, respectively. (**c**) Refractive index (*RI*) predicted using empirical formulas and prediction method of this study(*ERTR*$\mathcal{D}_4$). The solid black line indicates $x = y$.

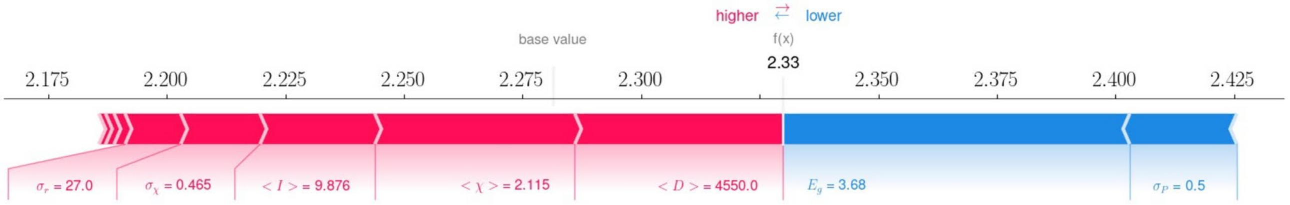

Fig. 6. SHAP force plot showing features impact on model prediction corresponding to *ZnS* which has the median *RI* value among unseen data (experimental *RI*: 2.35, prediction *RI*: 2.33).

with positive values increasing and negative values decreasing the prediction. Features are listed on the $y-$axis in order of importance. Each dot represents a data instance, colored from blue (low value) to red (high value). Taking advantage of the 10 features with the highest impact on the prediction of *RI*, we discuss other tree-based regression methods in the following section.

Concluding remarks

Given that the RFR method exhibits lower error compared to SVR and GPR, one might think that tree-based models outperforms other regression methods in prediction of *RI*. To investigate this, we perform extra analysis by using more complex and advanced tree-based ensemble models; namely the gradient boosted trees regression (GBTR) and extremely randomized trees regression (ERTR). We need to indicate that, GBTR combines the models by weighing them based on their performance, instead of taking the average of the predictions of all individual trees. This boosting is typically done through gradient descent optimization and thus, GBTR is more complex and computationally more expensive than RFR[40,41]. Also, ERTR builds multiple trees like RFR, but with even more randomization in the tree-building process, which makes it often faster to train compared to RFR[42].

Furthermore, considering that an increase in the number of features leads to overfiting for SVR and GPR, we evaluated model performance by increasing the number of predictors, one by one and according to SHAP analysis. Notably, the accuracy reaches its optimum when we keep 10 predictors and further increase in the number of predictors does not provide any higher accuracy. Accordingly, we select the 10 most influential features according to the SHAP analysis in Fig.5a. This data set is referred to as $\mathcal{D}_4$.

The results of *RI* prediction based on different tree-based methods by using $\mathcal{D}_i$ for $i \in \{1, 2, 3, 4\}$ is summarized in Fig. 5b. Here, we illustrate the MAE of the 10-fold cross-validation for RFR, GBTR, and ERTR. First, we notice that selecting $\mathcal{D}_4$ data set gives rise to a notably higher accuracy for all tree-based methods and reduces the computational cost. Therefore, we find it more reasonable to use $\mathcal{D}_4$ instead of $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$. Next, by concentrating on $\mathcal{D}_4$, we observe that more complex methods of GBTR and ERTR provide higher accuracy in comparison to RFR. To be more specific, the 10-fold cross-validation for the RFR, GBTR, and ERTR method with the $\mathcal{D}_4$ data set yield the MAE of 0.147, 0.143, and 0.135, respectively.

In Fig. 5c, we compare the *RI* prediction strength of our ML model for unseen data with empirical well-known relations. Here, we use ERTR since this model provides a more accurate prediction in comparison to other tree-based models. In this figure the horizontal axis represents the experimental *RI*, while the vertical axis stands for the predicted values. Consequently, the solid black line guides the eye for the best prediction and data points accuracy can be estimated by their distance to this line. Using the Ravindra formula to predict *RI* for unseen compounds, shows a large deviation from experimental values for $RI \in (1.75, 2.30)$, but it gives a reasonable prediction for $RI \in (2.5, 3.25)$. However, exploiting the trained ERTR model provides a much better prediction of refractive index for the whole range of *RI* in Fig. 5c. Also, Reddy, Herve, and Moss formulas give good predictions for $RI \in (1.75, 3.25)$, though our ML-assisted prediction still outperforms all these empirical formulas. Also, Table 1 lists the predicted *RI* values using ML models and empirical formulas which exhibits the least MAE and MAPE for ERTR using $\mathcal{D}_4$.

Focusing on a specific compound, in Fig. 6 we provide a SHAP force plot for *ZnS* which highlights the key features driving the model prediction, demonstrating their importance and directional impact. For this compound using ERTR, we observe a good prediction, namely $MAPE = 0.8\%$. As we see in this figure, the effect of E_g and σ_P in evaluating *RI* is shown by blue color (directed from left to right) indicating negative relationship and accordingly reducing the predicted *RI*. This is particularly consistent with the heatmap in Fig. 1b for E_g, as we observe large negative correlation. On the other hand, for $\langle D \rangle$, $\langle \chi \rangle$,$\langle I \rangle$,σ_χ and σ_r we observe a positive relationship (directed from right to left) indicated by red color, meaning that these predictive features increase the target variable *RI*.

To conclude, in this work, we investigate six different machine learning regression models, namely ordinary least square(OLSR), support vector (SVR), Gaussian process (GPR), random forest (RFR), gradient boosted trees (GBTR), and extremely randomized trees regression(ERTR) to predict the Refractive Index (RI) of diverse inorganic compounds. To scrutinize these models more closely, four data sets comprising 1, 5, 10, and 21 features are considered, which include properties such as band gap energy (E_g), electronegativity, first ionization energy, and other fundamental properties of the constituent elements of the inorganic compounds.

By comparing the predicted *RI*s with their experimental counterparts, we evaluate mean absolute error as a measure of accuracy in prediction for unseen compounds in the learning process. Our results reveal that although OLSR performance is poor in predicting *RI*s, other regression methods implemented in this work provide reasonably accurate estimation of *RI*s. Furthermore, we observe that increasing the number of predictors from 1 to 21 for SVR and GPR gives rise to the problem of model overfitting. In contrast, using tree-based methods with 10 predictors leads to a lower mean absolute error for unseen compounds. We have to emphasize that the accuracy of *RI* prediction using SVR and GPR, with band gap as the only predictor, is still higher than empirical formulas, namely Reddy, Ravindra, Moss, and Herve. Furthermore, utilizing ERTR with 10 predictors gives the highest accuracy in predicting *RI*s.

Overall, this research demonstrates that using machine learning for various inorganic compounds, particularly regression models, provides a reasonable tool for prediction of optical properties. Specially, ERTR is efficient in terms of processing time and numerical cost. In addition, the accuracy of machine learning-assisted models can be enhanced by considering a larger number of training compounds.

Data availibility

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 15 April 2024; Accepted: 18 September 2024
Published online: 15 October 2024

References

1. Hecht, E. *Optics* (Pearson Education India, 2012).
2. Weber, M. J. *Handbook of Optical Materials* Vol. 19 (CRC Press, 2002).
3. Korotkov, A. & Atuchin, V. Accurate prediction of refractive index of inorganic oxides by chemical formula. *J. Phys. Chem. Solids* **71**, 958–964 (2010).
4. Ghosh, D. & Samanta, L. Refractive indices of some narrow and wide bandgap materials. *Infrared Phys.* **26**, 335–336 (1986).

5. Bojan, M. et al. Refractive index measurement using comparative interferometry. In *Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies III*, Vol. 6635, 191–195 (SPIE, 2007).
6. Singh, S. Refractive index measurement and its applications. *Phys. Scr.* **65**, 167 (2002).
7. Lorentz, H. A. *Versuch einer Theorie der electrischen und optischen Erscheinungen in bewegten Körpern* (BG Teubner, 1906).
8. Kragh, H. The lorenz-lorentz formula: Origin and early history. *Substantia* **2**, 7–18 (2018).
9. Moss, T. Relations between the refractive index and energy gap of semiconductors. *Phys. Status Solidi (b)* **131**, 415–427 (1985).
10. Ravindra, N., Auluck, S. & Srivastava, V. On the penn gap in semiconductors. *Physica status solidi (b)* **93**, K155–K160 (1979).
11. Hervé, P. & Vandamme, L. General relation between refractive index and energy gap in semiconductors. *Infrared Phys. Technol.* **35**, 609–615 (1994).
12. Reddy, R., Ahammed, Y. N., Gopal, K. R. & Raghuram, D. Optical electronegativity and refractive index of materials. *Opt. Mater.* **10**, 95–100 (1998).
13. Ravindra, N., Ganapathy, P. & Choi, J. Energy gap-refractive index relations in semiconductors-an overview. *Infrared Phys. Technol.* **50**, 21–29 (2007).
14. Gomaa, H. M., Yahia, I. & Zahran, H. Correlation between the static refractive index and the optical bandgap: Review and new empirical approach. *Phys. B* **620**, 413246 (2021).
15. Duffy, J. Trends in energy gaps of binary compounds: an approach based upon electron transfer parameters from optical spectroscopy. *J. Phys. C: Solid State Phys.* **13**, 2979 (1980).
16. Mohan, S., Kato, E., Drennen, J. K. III. & Anderson, C. A. Refractive index measurement of pharmaceutical solids: a review of measurement methods and pharmaceutical applications. *J. Pharm. Sci.* **108**, 3478–3495 (2019).
17. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
18. Ramakrishna, S. et al. Materials informatics. *J. Intell. Manuf.* **30**, 2307–2326 (2019).
19. Noh, J., Gu, G. H., Kim, S. & Jung, Y. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* **11**, 4871–4881 (2020).
20. Audus, D. J. & de Pablo, J. J. Polymer informatics: Opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
21. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annu. Rev. Mater. Res.* **50**, 71–103 (2020).
22. Kang, X., Zhao, Y. & Li, J. Predicting refractive index of ionic liquids based on the extreme learning machine (elm) intelligence algorithm. *J. Mol. Liq.* **250**, 44–49 (2018).
23. Lightstone, J. P., Chen, L., Kim, C., Batra, R. & Ramprasad, R. Refractive index prediction models for polymers using machine learning. *J. Appl. Phys.* **127**, 215105 (2020).
24. Zhao, J. & Cole, J. M. Reconstructing chromatic-dispersion relations and predicting refractive indices using text mining and machine learning. *J. Chem. Inf. Model.* **62**, 2670–2684 (2022).
25. Zhou, Z.-H. *Machine Learning* (Springer Nature, 2021).
26. Carleo, G. et al. Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
27. Wei, J. et al. Machine learning in materials science. *InfoMat* **1**, 338–358 (2019).
28. Alzubi, J., Nayyar, A. & Kumar, A. Machine learning from theory to algorithms: an overview. In *Journal of Physics: Conference Series*, Vol. 1142, 012012 (IOP Publishing, 2018).
29. Liu, Z., Zhu, D., Raju, L. & Cai, W. Tackling photonic inverse design with machine learning. *Adv. Sci.* **8**, 2002923 (2021).
30. Mehta, P. et al. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).
31. Vapnik, V. Pattern recognition using generalized portrait method. *Autom. Remote. Control.* **24**, 774–780 (1963).
32. Zhang, F. & O'Donnell, L. J. Support vector regression. In *Machine learning*, 123–140 (Elsevier, 2020).
33. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
34. Scholkopf, B. & Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT Press, 2018).
35. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
36. Rasmussen, C. E. et al. *Gaussian Processes for Machine Learning* Vol. 1 (Springer, 2006).
37. Lizotte, D. J. et al. Automatic gait optimization with gaussian process regression. In *IJCAI*, Vol. **7**, 944–949 (2007).
38. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
39. Li, Y. et al. Random forest regression for online capacity estimation of lithium-ion batteries. *Appl. Energy* **232**, 197–210 (2018).
40. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).
41. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**, 21 (2013).
42. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
43. Breiman, L., Friedman, J., Stone, C. & Olshen, R. *Classification and Regression Trees* (Taylor & Francis, 1984).
44. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. & Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **71**, 804–818 (2015).
45. Singh, B., Sihag, P. & Singh, K. Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Model. Earth Syst. Environ.* **3**, 999–1004 (2017).
46. Zhou, X. et al. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J.* **4**, 212–219 (2016).
47. Grömping, U. Variable importance assessment in regression: linear regression versus random forest. *Am. Stat.* **63**, 308–319 (2009).
48. Haynes, W. M. *CRC Handbook of Chemistry and Physics* (CRC Press, 2016).
49. Zhao, X., Wang, X., Lin, H. & Wang, Z. Electronic polarizability and optical basicity of lanthanide oxides. *Phys. B* **392**, 132–136 (2007).
50. Wang, L., Tan, L., Hou, T. & Shi, J. Investigation of change regularity of energy states of mn2+ in halides. *J. Lumin.* **134**, 319–324 (2013).
51. Wang, L., Sun, Q., Liu, Q. & Shi, J. Investigation and application of quantitative relationship between sp energy levels of bi3+ ion and host lattice. *J. Solid State Chem.* **191**, 142–146 (2012).
52. Tripathy, S. K. & Pattanaik, A. Optical and electronic properties of some semiconductors from energy gaps. *Opt. Mater.* **53**, 123–133 (2016).
53. Reddy, R. et al. Interrelationship between structural, optical, electronic and elastic properties of materials. *J. Alloy. Compd.* **473**, 28–35 (2009).
54. Reddy, R. et al. Correlation between optical electronegativity and refractive index of ternary chalcopyrites, semiconductors, insulators, oxides and alkali halides. *Opt. Mater.* **31**, 209–212 (2008).
55. Reddy, R. et al. Correlation between optical electronegativity, molar refraction, ionicity and density of binary oxides, silicates and minerals. *Solid State Ionics* **176**, 401–407 (2005).
56. Reddy, R. et al. On the equivalence between clausius-mossotti and optical electronegativity relations. *Opt. Mater.* **22**, 7–11 (2003).
57. Reddy, R., Ahammed, Y. N., Gopal, K. R., Azeem, P. A. & Rao, T. Physico-chemical parameters of alkali halides using optical electronegativity. *Infrared Phys. Technol.* **42**, 49–54 (2001).

58. Phuoc, T. X., Wang, P. & McIntyre, D. Discovering the feasibility of using the radiation forces for recovering rare earth elements from coal power plant by-products. *Adv. Powder Technol.* **26**, 1465–1472 (2015).
59. Polyanskiy, M. N. Refractive index database. https://refractiveindex.info. (Accessed 11 Mar 2023).
60. Lee, J. S. & Lee, Y. H. Metal-to-metal antifuse with amorphous Ti-rich barium titanate film and silicon oxide film. *Solid-State Electron.* **43**, 469–472 (1999).
61. Korotkov, A. Correlation of optical properties of acentric crystals with chemical composition. *Opt. Commun.* **294**, 218–222 (2013).
62. Guo, Y. Y., Kuo, C. & Nicholson, P. S. The ionicity of binary oxides and silicates. *Solid State Ionics* **123**, 225–231 (1999).
63. Dickmann, J., Hurtado, C. R., Nawrodt, R. & Kroker, S. Influence of polarization and material on brownian thermal noise of binary grating reflectors. *Phys. Lett. A* **382**, 2275–2281 (2018).
64. Anani, M. et al. Model for calculating the refractive index of a iii–v semiconductor. *Comput. Mater. Sci.* **41**, 570–575 (2008).
65. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
66. Shannon, R. C., Lafuente, B., Shannon, R. D., Downs, R. T. & Fischer, R. X. Refractive indices of minerals and synthetic compounds. *Am. Miner.* **102**, 1906–1914 (2017).

Acknowledgements

We thank Joohwi Lee for sharing data and fruitful discussions.

Author contributions

M.M. conceived and supervised the project. E.E. performed the numerical calculations. All authors discussed the results and contributed to writing the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73551-0.

Correspondence and requests for materials should be addressed to M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.