The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Modeling the dielectric constants of crystals using machine learning

Cite as: J. Chem. Phys. 153, 024503 (2020); doi: 10.1063/5.0013136
Submitted: 8 May 2020 • Accepted: 22 June 2020 •
Published Online: 10 July 2020

Kazuki Morita,[1] Daniel W. Davies,[2] Keith T. Butler,[3,a)] and Aron Walsh[1,4,a)]

AFFILIATIONS

[1]Department of Materials, Imperial College London, London SW7 2AZ, United Kingdom
[2]Department of Chemistry, University College London, London WC1H 0AJ, United Kingdom
[3]SciML, Scientific Computer Division, Rutherford Appleton Laboratory, Harwell OX11 0QX, United Kingdom
[4]Department of Materials Science and Engineering, Yonsei University, Seoul 03722, South Korea

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.
[a)]**Author to whom correspondence should be addressed:** keith.butler@stfc.ac.uk and a.walsh@imperial.ac.uk

ABSTRACT

The relative permittivity of a crystal is a fundamental property that links microscopic chemical bonding to macroscopic electromagnetic response. Multiple models, including analytical, numerical, and statistical descriptions, have been made to understand and predict dielectric behavior. Analytical models are often limited to a particular type of compound, whereas machine learning (ML) models often lack interpretability. Here, we combine supervised ML, density functional perturbation theory, and analysis based on game theory to predict and explain the physical trends in optical dielectric constants of crystals. Two ML models, support vector regression and deep neural networks, were trained on a dataset of 1364 dielectric constants. Analysis of Shapley additive explanations of the ML models reveals that they recover correlations described by textbook Clausius–Mossotti and Penn models, which gives confidence in their ability to describe physical behavior, while providing superior predictive power.

I. INTRODUCTION

The dielectric response function is one of the fundamental properties of materials, which can give an insight into optical and electric properties. Specifically, the electronic high-frequency component of the dielectric constant has been explored intensively in early studies and is still of interest among researchers today.[1–3] Numerous efforts have been made to model dielectric constants. Most notably, the Clausius–Mossotti (CM) and Penn models, and their variants, are widely adopted throughout the literature.[2–5]

A. Clausius–Mossotti model

The Clausius–Mossotti (CM) equation expresses the dielectric constant ε as

$$\frac{\varepsilon - 1}{\varepsilon + 2} = \sum_i \frac{4\pi\alpha_i}{3v}. \tag{1}$$

Here, α_i is the polarizability of atomic species i and v is the volume of the unit cell. In the case of molecular crystals, α_i can be assigned to a constituent molecule,[6–8] and in ionic solids, it is assigned to an ion. The inherently many-body nature of the dielectric constant is reduced to this simple relation through the employment of two large assumptions.

First, an external electric field is screened by a dielectric medium before reaching an atom. In the CM model, cubic symmetry is assumed, and the local field $E^{\mathrm{loc}}(\boldsymbol{r})$ is expressed as

$$E^{\mathrm{loc}}(\boldsymbol{r}) = \frac{\varepsilon + 2}{3} E^{\mathrm{ext}}(\boldsymbol{r}), \tag{2}$$

where $E^{\mathrm{ext}}(\boldsymbol{r})$ is the external electric field. This relation, the Lorentz relation, is frequently used for non-cubic cases and generally holds when the anisotropy is small.[9,10] In the particular case where $E^{\mathrm{loc}}(\boldsymbol{r}) = E^{\mathrm{ext}}(\boldsymbol{r})$, the relation between the polarizability and the dielectric constant is called the Drude relation.[11]

Second, the atomic polarizability α_i is assumed to be a constant additive quantity. In other words, α_i is assumed to be unaffected by the environment. Due to its simple form, the CM model is still being used and, in practice, works for many materials,[2,12] despite the underlying assumptions.[9,13–15]

Numerous efforts have been made to improve the CM model.[16–18] In general, more effort was put into improving atomic polarizability α_i rather than changing the functional form.[16–19] While Wilson and Curtis increased the accuracy of the model by considering the electrostatic environment of an atom,[17] Jemmer *et al.* fitted parameters to Møller–Plesset calculations.[18] Nonetheless, the range of compounds and structures in which the model is valid was found to be limited. More recently, Shannon and Fischer fitted atomic polarizabilities to sets of experimental and computational data.[2,3] This was one of the first studies to consider compounds throughout a wider variety of structures, and their values of atomic polarizability were more transferable. Although many studies focus on materials that are well described by a particular model, Shannon and Fischer investigated materials that could not be described by their model. This approach allowed them to identify anomalous dielectric behavior in materials that are sterically strained or have the perovskite structure.

B. Penn model

The premise of the Penn model is the electronic band picture. When the overlap between electrons at different atomic sites is large, such a model is anticipated to be favorable. Treating the external electric field as a time-dependent perturbation, Penn derived the following relation:[1]

$$\varepsilon \approx 1 + \left(\frac{\hbar\omega_p}{\mathscr{E}_g}\right)^2, \tag{3}$$

where ω_p is the plasma frequency and $\mathscr{E}_g$ is the width of the bandgap. The significance lies in the fact that the dielectric constant is described only by the plasma frequency and the bandgap. Using the second order perturbation theory, polarizability α for an isolated atom can be written as $\alpha \approx (\hbar\omega_p/\mathscr{E}_g)^2$, where $\mathscr{E}_g$ is the HOMO–LUMO gap. In this case, the screening of the external field can be ignored, and the Penn model is equivalent to the CM model.

As with the CM model, there are various assumptions for the Penn model. The first main assumption is to consider the electrons as a free electron gas. This allows the form of the equation to be greatly simplified but is not valid when electrons are localized.[20,21] The second assumption is to consider only the bandgap and not the density of states around the valence and conduction band edges. This flat band approximation results in a large error when the valence or conduction band has a large dispersion.[5]

Various efforts have been made to improve this model. Phillips added a correction for ionic bonds and showed that dielectric constants in zinc blende and wurtzite were reproduced.[20,21] Furthermore, Van Vechten added a scaling factor taking into account that not all valence electrons contribute to the dielectric response.[22,23] These modifications did contribute to increasing the predictive power; however, a universal model was not achieved.

In addition to CM/Penn based models, many other empirical models such as those of Allen, Gladstone–Dale (GD), and Anderson–Eggleton have been developed.[24–27] All of these models relate atomic polarizability α to dielectric constant ε. Within these models, the GD model has been researched intensively;[3,28–30] however, Fowler *et al.* reported that the GD model overestimates the atomic polarizabilities and the CM model is theoretically more rigorous.[29]

C. Data driven models

Rapid progress in computer technology and first-principles modeling techniques has enabled the calculation of dielectric constants for a large number of materials.[31–33] The modern calculation of dielectric constants is typically done using density functional perturbation theory (DFPT).[4,34–38] Despite being a perturbation theory, DFPT can be calculated with a relatively low cost, and the results are less dependent on the exchange–correlation functional used, compared to properties such as bandgaps.[4] Therefore, many studies have involved analysis of datasets of dielectric constants aiming to understand trends or predict values for new materials. For example, Han and co-workers performed high-throughput calculation of binary and ternary inorganic compounds and compared the relation between the bandgap and the dielectric constant.[4,34] Some studies employ machine learning (ML) methods where a statistical model is trained.[35,36] For example, Umeda *et al.* trained different ML models on a dataset of 3382 compounds.[35] They obtained good agreement with DFPT calculations; however, the reasons for this agreement are not discussed.

The CM model approaches the problem from a molecular picture, whereas the Penn model is based on the electronic band picture of a crystal. As such, the former has difficulty with delocalized electrons, and the latter has difficulty with localized electrons. In both cases, the improvement of the model was done empirically, and when the number of parameters became large, systematic expansion of the model became difficult. On the other hand, ML models can increase model complexity systematically, while they suffer from poor interpretability. By studying the difference between ML and analytical models, we can recover new underlying physics and deepen our understanding of the phenomenological behavior of the dielectric response.

In this study, we use two types of ML models and investigate the reasoning behind their predictions. First, a deep neural network (DNN) and a support vector regression (SVR) model are trained on the same dataset. Next, their results are compared with DFPT calculations. Finally, we perform game theoretic analysis [Shapley additive explanations (SHAP)] to elucidate the characteristics of the trained ML models and to investigate the reason behind their predictions. We compare the result closely with the CM and Penn models and show that ML can indeed learn the underlying physical trends while having superior predictive power. Additionally, the Shapley analysis allows us to identify reasons for poor model performance in particular cases, which is important when considering how reliable the predictions are.

II. METHODOLOGY

A. Machine learning

Supervised ML models were trained on a dataset of 1364 dielectric constants. The dataset was prepared by combining two

pre-existing datasets[37,38] and averaging the data for the overlapping materials. Materials with a bandgap less than 0.5 eV were removed because small gap materials require very dense sampling of the Brillouin zone, which is difficult to realize in high-throughput calculations.[43] The dielectric constant was calculated by taking a diagonal average of the electronic part of the dielectric matrix. Since this dataset only contained minimum features, it was augmented and processed using *Materials Project*,[39,40] *pymatgen*,[42] *SMACT*,[41] and *scikit-learn*.[44] The dataset was split into 8:2, the training:test sets. The same dataset was used when training the two different ML models.

The first ML model was support vector regression (SVR). Before training, we analyzed the available features and removed those unimportant or similar. This step was necessary because having too many features increases the dimensions of the manifold that the model must learn in and thus lowers the performance.[45] Specifically, we calculated feature importance using the random forest (Fig. S1) and removed low importance features: space group (one hot encoded), atomic species (one hot encoded), number of elements, and number of sites inside the unit cell. Next, with the criterion of $r^2 > 0.90$, we removed the maximum oxidation state and variation in Madelung energy, which had high correlation with variation in oxidation state and minimum in Madelung potential, respectively (Fig. S2). The refined features are presented in Table I.

Four machine learning models, the random forest, gradient boost regression, kernel ridge regression, and SVR, were trained and compared (Fig. S4 and Table S1). The former two models are based on decision trees, and the latter two models are based on kernel methods. Since the random forest and gradient boost regression are ensemble methods, we surmise their performance to be better; however, they were overfitting, and the kernel methods were able to generalize better. The kernel ridge regression had a slightly lower performance compared to SVR; therefore, SVR was employed for this study.

The second ML model was a deep neural network (DNN). Specifically, we have adopted the DNN architecture *MEGNet* developed by Chen *et al.*[46] *MEGNet* can be trained only by using the crystal structure of materials. The exact structure of the network is presented in Fig. S3. It is not trivial to uniquely express a crystalline system in the form of a vector; however, *MEGNet* overcomes this difficultly by representing bonding networks as graphs and using the set2set algorithm to consistently treat different sized graphs.[46] Atomic number and bond length are encoded into the graph representation. The model was trained for 400 epochs, and the layer weights of the epoch with the smallest Huber loss were employed.

There are two advantages and a disadvantage of using *MEGNet* over SVR. The first advantage is to avoid the so-called "feature engineering." Feature engineering is a process of feature selection, which was necessary for SVR. Although this can be a way of inputting our domain knowledge,[31] this relies heavily on experience and intuition and may obstruct the systematic improvement of an ML model. The second advantage is that it is easy to perform transfer learning. As larger datasets are available for other materials properties, it is tempting to use that information in training. Since the upper layers in DNNs are known to learn general trends, it is possible to import the layer weights of the upper layers in advance and improve the accuracy.[46] Taking into account the Penn model, and the results of feature engineering for SVR (Figs. S2 and S1), we performed transfer learning from the *MEGNet* model that Chen *et al.* trained on bandgaps.[46] The disadvantage of DNNs is interpretability. Specifically, because we used graph representation as the input, the relation between the features and dielectric constant were difficult to extract.

TABLE I. Features used to train the support vector regression model.

Feature	Dimensions
Bandgap[a]	1
Δ Pauling energy[b]	1
Material density[a]	1
Formation energy (per atom)[a]	1
Oxidation state (minimum, variation)[a]	2
Madelung energy (minimum, maximum)[c]	2
Ionic species (one hot encoded)[c]	85

[a]Obtained from the Materials Project.[39,40]
[b]Calculated using SMACT.[41]
[c]Calculated using the pymatgen package.[42]

B. Density functional theory calculation

Following the approach of Shannon and Fischer,[2,3] we have investigated materials that had large error in our SVR and DNN predictions. To do this, we calculated the total of 24 structures using DFPT. The calculations were performed with the projector-augmented wave scheme as implemented in VASP.[47–49] The structures were taken from *Materials Project* and were calculated using the PBE functional.[39,40,50] The reciprocal space was sampled so that the spacing between the k-point was about $2\pi \times 0.03$ Å^{-1}. The energy cutoff was set to at least 600 eV, and the wavefunctions were optimized to a tolerance of 10^{-6} eV.

C. Shapley additive explanations

For model interrogation, we perform Shapley additive explanations (SHAP) analysis.[51] The Shapley regression value is defined as

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]. \quad (4)$$

Here, S is a subset of the features (F), i is a particular feature of interest, f is the ML model, and x_S represent the values of the input features in the set S.[51] ϕ_i describes how much the model output changes when feature i is added to the model. Therefore, it can be used to quantify feature importance. As SVR cannot take features of different sizes, we used the median of the feature instead of removing them. Furthermore, if we fix data x_S and model f, we can show which features are responsible for the prediction given subset x_S. Using the additivity approximation suggested by Lundberg and Lee,[51] we calculated the SHAP values of all the features and data in the dataset.

III. RESULT

A. ML prediction

The dielectric constants predicted by SVR are plotted against the dataset values in Fig. 1(a). In general, materials with dielectric constants above 10 had larger error. The DNN performed similarly to SVR, as shown in Fig. 1(b). The main difference was suppression of error in the high dielectric constant range and a slight increase of error in the low dielectric constant range in the predictions from the DNN.

The performance metrics of SVR and DNN are summarized in Table II. This performance is similar to the performance of ML models in previous studies.[52–54] Ideally, if there is no overfitting and the ML model is trained sufficiently, metrics for training and test data should be identical; however, Table II implies tendency for both models to exhibit overfitting. For SVR, different sets of hyperparameters and loss functions were trialed, and the result presented here represents the best achievable performance (Table S1). We built a series of SVR models on different sized subsets of the data (Fig. S6) and find that the difference in the loss function metric decreased as we increased the dataset size but did not converge. We therefore attribute overfitting to the small size of our dataset. Compared to SVR, the difference in metric between the test and training data was larger for the DNN, suggesting that the former model was able to generalize better and exhibits overfitting to a lesser extent. For test data, the SVR demonstrated higher performance for all of the metrics than the DNN (Table II).

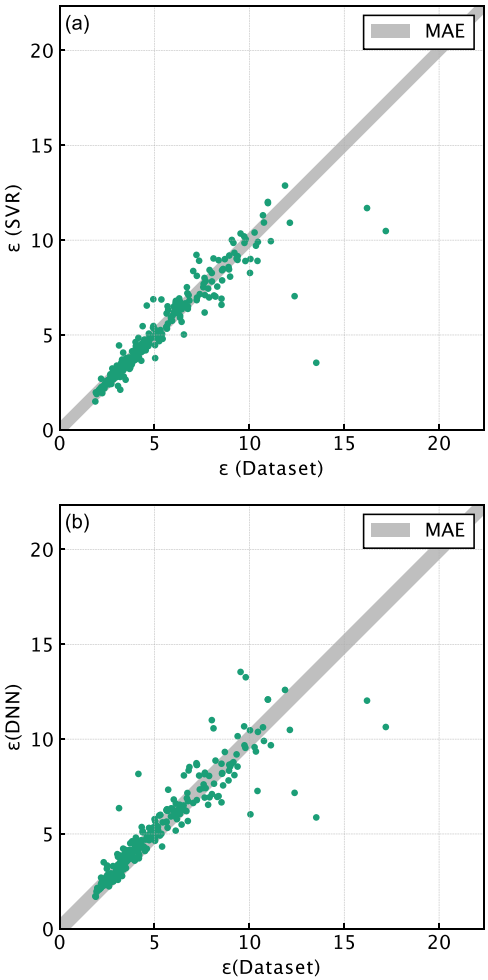

FIG. 1. Prediction of (a) support vector machine (SVR) and (b) deep neural network (DNN) compared with dataset values. Mean average error (MAE) of the ML prediction is shaded gray.

TABLE II. Performance metrics of support vector regression (SVR) and deep neural network (DNN) for training and test data. Metrics are mean Pearson's correlation coefficient (r^2), average error (MAE), mean squared error (MSE), and root mean square error (RMSE).

	SVR		DNN	
Metric	Training	Test	Training	Test
r^2	0.92	0.86	0.95	0.84
MAE	0.24	0.44	0.20	0.55
MSE	0.69	0.99	0.38	1.17
RMSE	0.83	0.99	0.62	1.08

The ten materials with largest prediction error are listed in Tables III and IV. Out of the ten materials, seven of them were common between the two ML methods. It is worth noting that LiBC, Ga_2Te_5, $LiAsS_2$, and In_2HgTe_4 have error large enough to be identifiable in Fig. 1. We also present the energy above hull (obtained from *Materials Project*) to show that these materials are not artifacts from computational screening and could possibly be stable.[39,40]

Since the dielectric constant dataset we use for training was derived from a high-throughput workflow,[37,38] the precision is optimized to be convergent for most materials,[55] which results in some

TABLE III. List of ten materials with largest error in support vector regression (SVR) prediction. ε_{SVR}, $\varepsilon_{Dataset}$, and ε_{DFPT} (this work) are dielectric constants from SVR prediction, training dataset, and our density functional perturbation theory (DFPT) calculation, respectively. $\Delta\varepsilon_{dataset}$ is the difference between ε_{SVR} and $\varepsilon_{Dataset}$. $\Delta\varepsilon_{DFPT}$ is the difference between ε_{SVR} and ε_{DFPT}. ΔE_{hull} (meV) is the energy above hull per atom.

Formula	ε_{SVR}	$\varepsilon_{Dataset}$	$\Delta\varepsilon_{dataset}$	ε_{DFPT} (this work)	$\Delta\varepsilon_{DFPT}$	ΔE_{hull} (meV)
LiBC	3.54	13.53	−9.99	11.96	−8.43	0.00
Ga_2Te_5	10.48	17.19	−6.71	13.75	−3.27	0.00
$LiAsS_2$	7.05	12.39	−5.34	8.57	−1.53	0.00
In_2HgTe_4	11.69	16.20	−4.51	13.30	−1.61	0.00
KCaBi	9.22	7.21	2.01	7.51	1.71	0.00
$CdCN_2$	6.55	4.59	1.96	4.45	2.10	0.00
HfNCl	6.88	4.94	1.95	4.68	2.21	0.00
$CuBS_2$	6.59	8.53	−1.94	8.31	−1.71	0.00
$CuBSe_2$	8.27	10.03	−1.76	9.78	−1.51	0.00
Cu_2HgI_4	6.91	8.53	−1.62	7.63	−0.72	0.00

TABLE IV. List of ten materials with largest error in deep neural network (DNN) prediction. ε_{DNN}, $\varepsilon_{Dataset}$, and ε_{DFPT} (this work) are dielectric constants from DNN prediction, training dataset, and our density functional perturbation theory (DFPT) calculation, respectively. $\Delta\varepsilon_{dataset}$ is the difference between ε_{DNN} and $\varepsilon_{Dataset}$. $\Delta\varepsilon_{DFPT}$ is the difference between ε_{DNN} and ε_{DFPT}. ΔE_{hull} (meV) is the energy above hull per atom.

Formula	ε_{DNN}	$\varepsilon_{Dataset}$	$\Delta\varepsilon_{dataset}$	ε_{DFPT} (this work)	$\Delta\varepsilon_{DFPT}$	ΔE_{hull} (meV)
LiBC	5.87	13.53	−7.66	11.96	−6.09	0.00
Ga_2Te_5	10.64	17.19	−6.55	13.75	−3.11	0.00
$LiAsS_2$	7.17	12.39	−5.22	8.57	−1.41	0.00
In_2HgTe_4	12.03	16.20	−4.18	13.30	−1.27	0.00
LiZnN	6.03	10.06	−4.02	9.85	−3.82	0.00
Cs_2HfI_6	8.17	4.15	4.01	3.81	4.35	0.00
AlAs[a]	13.54	9.54	4.00	9.59	3.95	6.27
AlAs[b]	13.26	9.81	3.45	9.72	3.54	0.00
Li_3NbS_4	6.36	3.13	3.23	3.03	3.33	6.61
$MgTe_2$	7.27	10.43	−3.16	10.42	−3.15	7.17

[a]Wurtzite structure.
[b]Zinc blende structure.

results being poorly converged.[43] Therefore, for comparison, we have performed higher precision calculations using DFPT on these compounds, and the results are presented in Tables III and IV.

IV. DISCUSSION

A. ML prediction

The larger error found when the large dielectric constant is greater is not surprising (Fig. 1). In addition to the numerical instability in DFPT calculations,[43] there are fewer materials with large dielectric constants in the dataset; therefore, the ML models were not able to fully learn the trends in the large dielectric constant range.

Comparing SVR and DNN, the performance was similar, as shown in Fig. 1 and Table II. This is surprising since the features used for training each model were different. It suggests that both models are sufficiently capturing physical trends in the training data. If the model is generalized well, we expect them to be able to detect anomalous data in the DFPT dataset. For eight (seven) materials out of ten materials, SVR (DNN) actually predicts values closer to our calculations, which were performed under higher precision. This is clearly the case for Ga_2Te_5 and $LiAsS_2$. For example, in the case of Ga_2Te_5 in Table III, the SVR prediction was 10.48, whereas the dataset value was 17.19 and our calculation was 13.75.

Since small bandgap materials require fine Brillouin zone sampling,[43] the standardized sampling in the high-throughput calculation setup may be insufficient. To confirm this, we calculated the dielectric constant of $LiAsS_2$ with different Brillouin zone sampling densities. Our converged value of dielectric constant was 8.57, while 11.46 was obtained using coarse sampling. 11.46 is closer to the dataset value of 12.39 (Table III and Table S5). Although this is not direct evidence, we speculate that the materials which exhibited a large difference between our calculations and the dataset were especially sensitive to the Brillouin zone sampling and, as a result, made the reported dielectric constants of these materials anomalous.

Although we found few cases of under-converged data, we believe that the overall precision of the dataset was sufficient. Since the training error serves as an upper bound of the random noise in the training dataset,[56] we can estimate that the error due to insufficient Brillouin zone sampling is smaller than the mean average error of 0.44 (SVR-test column in Table II). Therefore, materials in Tables III and IV have significantly larger numerical error than the average. This trend is also apparent when compared with the materials with small prediction error (Tables S3 and S4).

B. SHAP analysis

SHAP analysis applies a game theoretic approach to calculate the importance of individual input features to a given model prediction. A positive (negative) SHAP value indicates that a given feature contributes to an increase (decrease) in the prediction with respect to the mean of the set. Figure 2 shows the calculated SHAP values for all features and data. The features are ordered by their importance. Note that the order of importance differs from Fig. S1, which was a random forest model trained on all features and data, whereas Fig. 2 is SVR built only on training data.

The high contribution of material density can be explained by both the CM and Penn expressions or explicitly related by the Gladstone–Dale and Allen models.[24,25] If there are more electrons in a given volume, the dielectric response will become larger, and indeed, SHAP analysis shows that dielectric constant monotonically increases with density.

The high contribution of the bandgap is also no surprise from the Penn model [Eq. (3)]. Lower energy excitations result in a larger dielectric constant. A large bandgap gives a negative SHAP

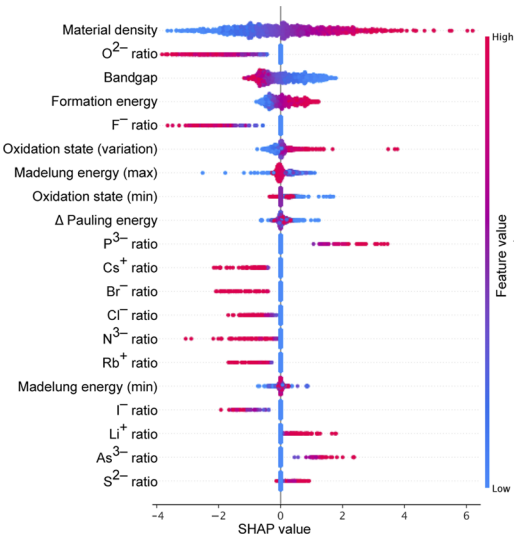

FIG. 2. Relation of the value of each features and their SHAP values. Plot is colored red (blue) if the value of the feature is high (low). The vertical width of the plot shows the number of points within the same SHAP value.

contribution. Interestingly, the magnitude of positive contributions from small bandgaps has a longer tail in the distribution than the negative contribution from large gaps. This suggests that although a large gap decreases the dielectric response, this effect diminishes quickly so that continuously increasing bandgap will not always decrease the dielectric constant.

To take a deeper look at these relationships, further analysis was performed, as shown in Fig. 3(a). The contribution from the bandgap decreases as the gap increases with an inverse power law relation. By coloring the individual points according to material density, the interplay of density and bandgap can be observed. Specifically, when the bandgap is low, a low density increases the SHAP value, while when the gap is high a low density decreases the magnitude of the negative SHAP value. While it is not possible to derive rigorous analytical relationship from the number of data points we have, Fig. 3(a) suggests a $\phi_{\mathscr{E}_g} \sim n/\mathscr{E}_g^m$ relation, where $\phi_{\mathscr{E}_g}$ is the SHAP value for the bandgap, n is the material density, and m is an arbitrary constant. This is similar to the Penn relation [Eq. (3)]. Therefore, it is possible to interpret that ML is learning the Penn model while incorporating other feature relations as correction terms.

When the formation energy is high (low), the dielectric constant is raised (suppressed) (see Fig. 2). This relation is opposite to that of the bandgap, and the negative correlation −0.65 between the two also agrees with this trend (Fig. S2). Within our ML model, two roles of formation energy can be suggested. First, it is acting as an ionicity parameter. According to Pauling, the formation energy of binary compounds is $\Delta H \propto (\Delta E_{\text{Pauling}})^2$, where $\Delta E_{\text{Pauling}}$ is the Pauling energy difference.[57] This trend was present in our dataset (Fig. S7); however, the variance was large, which suggests contributions from other features. Second, it is another feature for the bandgap. In main group binary compounds, the relation $\Delta H \propto (\mathscr{E}_g^2/\mathscr{E}_f)\ln(\mathscr{E}_g/\mathscr{E}_f)$ has been reported.[58] Quantitatively, this relation did not hold for our case (Fig. S7). Given these two relations and accounting for the fact that formation energy did not show an obvious relation between the Pauling energy difference and the bandgap, we suggest that it contains both information weakly.

The variation of oxidation states also exhibited high importance (Fig. 2). Since oxidation states contain information about ionicity, we may expect the variation in oxidation state to lower the value of the dielectric constant when it is large.[22,23] Figure 2 shows the

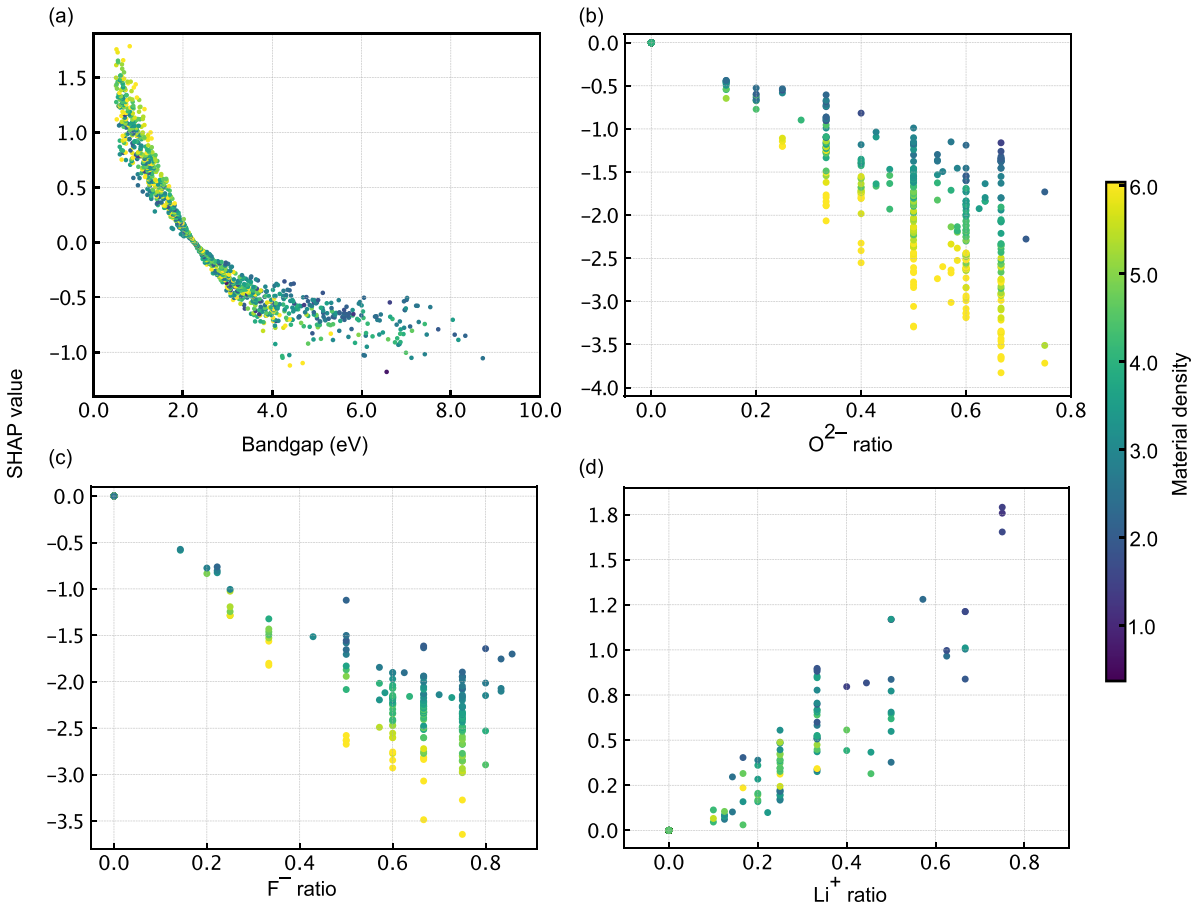

FIG. 3. SHAP values for (a) bandgap, (b) O^{2-} ratio, (c) F^- ratio, and (d) Li^+ ratio. The points are colored according to material density.

opposite trend. This counter intuitive result can be explained by interpreting this feature as a correction to the flat band approximation, described earlier for the Penn model. As the bandgap is defined as a difference between the valence band maximum and the conduction band minimum, it does not have information about band dispersion. If ionic and covalent compounds with the same bandgap exist, ionic compounds will have smaller effective bandgaps. When the bandgap is large (small) and contributes to making dielectric constant smaller (larger), lower variation in oxidation states amplifies (suppresses) this effect (Fig. S8). Theoretical studies also support the importance of this correction.[5]

Other oxidation state and Madelung energy features did not show clear trends, which constitutes further evidence that the variation of oxidation state is acting as an ionicity parameter. This result would not have been available if we were only considering feature importance and highlights the ability of SHAP to access a deeper understanding of trained models.

Since the distribution of ionic species ratio features had broad SHAP value distributions and were difficult to interpret in Fig. 2, we continue their discussion based on Figs. 3(b)–3(d). The SHAP values of the O^{2-} and F^{-} ratio are plotted against material density. In general, higher concentration of these ionic species tends to reduce the dielectric constant. This effect could be interpreted as originating from the strongly electronegative nature of O and F, which stabilize valence electrons and suppress their dielectric response. In contrast, for the Li^{+} cation, the SHAP value is positive, meaning that it increases the dielectric constant. Furthermore, we can see the interplay of density and composition; when the density is higher, the effect of the O^{2-} and F^{-} ratio was enhanced, where O^{2-} had a slightly larger change. Due to the limited number of data points, further study is required to draw firmer conclusions. Furthermore, it should be noted that the SHAP values for ionic species cannot be directly related to atomic polarizabilities because our model is not explicitly based on physical models such as the CM model.

Finally, we explore the limitations of the trained SVR model by examining LiBC and $CdCN_2$ in Table III. Since SHAP values are additive,[51] we are able to decompose them to each of the features. We begin with the dielectric constant predicted for the median of features ($\varepsilon = 5.58$) and add contribution of all the features sequentially. Although only the ten most important features are presented, other features had minor contribution ~0.001, so we will only discuss features presented in Fig. 4. For LiBC, which had largest error for both SVR and DNN, SVR predicted 3.54 where our calculated value was 11.96 (Table III). As shown in Fig. 4(a), the largest contribution was from the material density. Together with the fact that the material density was small (2.15 g/cm^2), the error is likely due to the incorrect contribution of material density. It is likely that in this range of material density, sufficient sampling was not achieved with our training dataset.

Most of the materials with large error feature a large contribution from material density. The exception was $CdCN_2$. SVR predicted 6.55, where 4.45 was the result of our DFPT calculation (Table III). As shown in Fig. 4(b), the three largest contributions are N^{3-} ratio, formation energy, and variation in oxidation state, which have reasonable values. We instead assign the error as arising from the large anisotropy, which is not taken into account. The full dielectric matrix is

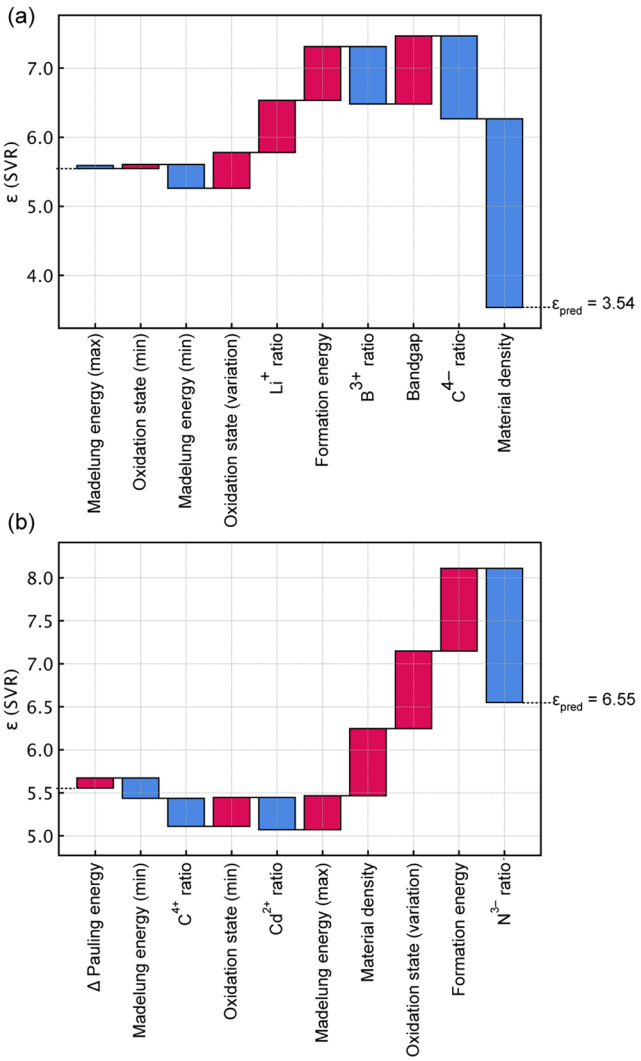

FIG. 4. Breakdown of SHAP values for ten most important features shown as a waterfall plot for (a) LiBC and (b) $CdCN_2$ in Table III. Starting with the dielectric constant predicted by a median of the features ($\varepsilon = 5.58$), each feature contributes to lowering or elevating the value of the dielectric constant in an additive manner. The final value of the dielectric constant ε_{pred} is the output of the model.

$$\varepsilon = \begin{bmatrix} 3.40 & 0.00 & 0.00 \\ 0.00 & 3.40 & 0.00 \\ 0.00 & 0.00 & 6.55 \end{bmatrix}. \quad (5)$$

The DNN performed better because the bond lengths are explicitly taken into account through the graph representation.

V. CONCLUSIONS

We showed that two different machine learning models, a support vector machine and a deep neural network, were able to predict the dielectric constants of crystals with reasonable precision.

Comparison with our density functional perturbation theory calculations reveals that the machine learning models were able to detect erroneous results in the original dataset. We performed the SHAP analysis of the support vector machine model, which illustrated that it is learning similar relations to the textbook Clausius–Mossotti and Penn expressions. Finally, we showed the limitation of our support vector model through a detailed analysis of the predictions.

We suggest that as long as the dataset is sufficiently large to sample the crystal space of interest, machine learning models can be an effective approach not only to predict material properties but also to capture physical trends. The analysis approach used in this study is not restricted to dielectric response and therefore has application potential for other relations, including properties that are not as intensively studied and where there is no existing analytical description.

SUPPLEMENTARY MATERIAL

See the supplementary material for additional model parameters and analysis (pdf). In addition, a set of electronic notebooks to reproduce the model training and analysis performed in this study.

ACKNOWLEDGMENTS

The authors thank financial support from the Yoshida Scholarship Foundation, the Japan Student Services Organization, and the Centre for Doctoral Training on Theory and Simulation of Materials at Imperial College London. This research was also supported by the Creative Materials Discovery Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (Grant No. 2018M3D1A1058536). Through our membership of the UK's HEC Materials Chemistry Consortium, which is funded by the EPSRC (Grant No. EP/L000202), this work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk).

REFERENCES

[1] D. R. Penn, "Wave-number-dependent dielectric function of semiconductors," Phys. Rev. **128**, 2093 (1962).
[2] R. D. Shannon and R. X. Fischer, "Empirical electronic polarizabilities in oxides, hydroxides, oxyfluorides, and oxychlorides," Phys. Rev. B **73**, 235111 (2006).
[3] R. D. Shannon and R. X. Fischer, "Empirical electronic polarizabilities of ions for the prediction and interpretation of refractive indices: Oxides and oxysalts," Am. Mineral. **101**, 2288–2300 (2016).
[4] K. Yim, Y. Yong, J. Lee, K. Lee, H.-H. Nahm, J. Yoo, C. Lee, C. Seong Hwang, and S. Han, "Novel high-κ dielectrics for next-generation electronic devices screened by automated *ab initio* calculations," NPG Asia Mater. **7**, e190 (2015).
[5] F. Naccarato, F. Ricci, J. Suntivich, G. Hautier, L. Wirtz, and G.-M. Rignanese, "Searching for materials with high refractive index and wide band gap: A first-principles high-throughput study," Phys. Rev. Mater. **3**, 044602 (2019).
[6] T. Luty, "On the effective molecular polarizability in molecular crystals," Chem. Phys. Lett. **44**, 335–338 (1976).
[7] K. J. Miller and J. Savchik, "A new empirical method to calculate average molecular polarizabilities," J. Am. Chem. Soc. **101**, 7206–7213 (1979).
[8] K. J. Miller, "Calculation of the molecular polarizability tensor," J. Am. Chem. Soc. **112**, 8543–8551 (1990).
[9] J. R. Tessman, A. H. Kahn, and W. Shockley, "Electronic polarizabilities of ions in crystals," Phys. Rev. **92**, 890 (1953).
[10] K. Urano and M. Inoue, "Clausius–Mossotti formula for anisotropic dielectrics," J. Chem. Phys. **66**, 791–794 (1977).
[11] N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals*, 2nd ed. (Oxford University Press, London, 1950).
[12] A. S. Korotkov and V. V. Atuchin, "Prediction of refractive index of inorganic compound by chemical formula," Opt. Commun. **281**, 2132–2138 (2008).
[13] S. T. Pantelides, "Mechanisms that determine the electronic dielectric constants of ionic crystals," Phys. Rev. Lett. **35**, 250 (1975).
[14] H. Coker, "Empirical free-ion polarizabilities of the alkali metal, alkaline earth metal, and halide ions," J. Phys. Chem. **80**, 2078–2084 (1976).
[15] P. W. Fowler and N. C. Pyper, "In-crystal ionic polarizabilities derived by combining experimental and *ab initio* results," Proc. R. Soc. London, A **398**, 377–393 (1985).
[16] A. R. Ruffa, "Theory of the electronic polarizabilities of ions in crystals: Application to the alkali halide crystals," Phys. Rev. **130**, 1412 (1963).
[17] J. N. Wilson and R. M. Curtis, "Dipole polarizabilities of ions in alkali halide crystals," J. Phys. Chem. **74**, 187–196 (1970).
[18] P. Jemmer, P. W. Fowler, M. Wilson, and P. A. Madden, "Environmental effects on anion polarizability: Variation with lattice parameter and coordination number," J. Phys. Chem. A **102**, 8377–8385 (1998).
[19] V. Dimitrov and S. Sakka, "Electronic oxide polarizability and optical basicity of simple oxides. I," J. Appl. Phys. **79**, 1736–1740 (1996).
[20] J. C. Phillips, "A posteriori theory of covalent bonding," Phys. Rev. Lett. **19**, 415 (1967).
[21] J. C. Phillips, "Dielectric definition of electronegativity," Phys. Rev. Lett. **20**, 550 (1968).
[22] J. A. Van Vechten, "Quantum dielectric theory of electronegativity in covalent systems. I. Electronic dielectric constant," Phys. Rev. **182**, 891 (1969).
[23] J. A. Van Vechten, "Quantum dielectric theory of electronegativity in covalent systems. II. Ionization potentials and interband transition energies," Phys. Rev. **187**, 1007 (1969).
[24] J. H. Gladstone and T. P. Dale, "XIV. Researches on the refraction, dispersion, and sensitiveness of liquids," Philos. Trans. R. Soc. London **153**, 317–343 (1863).
[25] R. D. Allen, "A new equation relating index of refraction and specific gravity," Am. Mineral. **41**, 245–257 (1956).
[26] O. L. Anderson and E. Schreiber, "The relation between refractive index and density of minerals related to the earth's mantle," J. Geophys. Res. **70**, 1463–1471, https://doi.org/10.1029/jz070i006p01463 (1965).
[27] R. A. Eggleton, "Gladstone-dale constants for the major elements in silicates; coordination number, polarizability, and the Lorentz–Lorentz relation," Can. Mineral. **29**, 525–532 (1991).
[28] F. D. Bloss, M. Gunter, S.-C. Su, and H. Wolfe, "Gladstone-dale constants; a new approach," Can. Mineral. **21**, 93–99 (1983).
[29] P. W. Fowler, R. W. Munn, and P. Tole, "Polarisability of the oxide ion in crystalline BeO," Chem. Phys. Lett. **176**, 439–445 (1991).
[30] J. A. Mandarino, "The Gladstone Dale compatibility of minerals and its use in selecting mineral species for further study," Can. Mineral. **45**, 1307–1324 (2007).
[31] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).
[32] J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer *et al.*, "New frontiers for the materials genome initiative," npj Comput. Mater. **5**, 41 (2019).
[33] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm," arXiv:2005.00707 (2020).
[34] M. Lee, Y. Youn, K. Yim, and S. Han, "High-throughput *ab initio* calculations on dielectric constant and band gap of non-oxide dielectrics," Sci. Rep. **8**, 14794 (2018).

[35]Y. Umeda, H. Hayashi, H. Moriwake, and I. Tanaka, "Prediction of dielectric constants using a combination of first principles calculations and machine learning," Jpn. J. Appl. Phys., Part 2 **58**, SLLC01 (2019).
[36]Y. Noda, M. Otake, and M. Nakayama, "Descriptors for dielectric constants of perovskite-type oxides by materials informatics with first-principles density functional theory," Sci. Techol. Adv. Mater. **21**, 92–99 (2020).
[37]I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T. D. Schladt, K. A. Persson, and F. B. Prinz, "High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials," Sci. Data **4**, 160134 (2017).
[38]G. Petretto, S. Dwaraknath, H. P. Miranda, D. Winston, M. Giantomassi, M. J. Van Setten, X. Gonze, K. A. Persson, G. Hautier, and G.-M. Rignanese, "High-throughput density-functional perturbation theory phonons for inorganic materials," Sci. Data **5**, 180065 (2018).
[39]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[40]S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, "The materials application programming interface (API): A simple, flexible and efficient API for materials data based on representational state transfer (REST) principles," Comput. Mater. Sci. **97**, 209–215 (2015).
[41]D. Davies, K. Butler, A. Jackson, J. Skelton, K. Morita, and A. Walsh, "SMACT: Semiconducting materials by analogy and chemical theory," J. Open Source Software **4**, 1361 (2019).
[42]S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).
[43]S. Baroni, S. De Gironcoli, A. Dal Corso, and P. Giannozzi, "Phonons and related crystal properties from density-functional perturbation theory," Rev. Mod. Phys. **73**, 515 (2001).
[44]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[45]A. E. Teschendorff, "Avoiding common pitfalls in machine learning omic data science," Nat. Mater. **18**, 422–427 (2018).
[46]C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," Chem. Mater. **31**, 3564–3572 (2019).
[47]P. E. Blöchl, "Projector augmented-wave method," Phys. Rev. B **50**, 17953 (1994).
[48]G. Kresse and J. Furthmüller, "Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6**, 15–50 (1996).
[49]G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169 (1996).
[50]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).
[51]S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 4765–4774.
[52]A. Mannodi-Kanakkithodi, G. Pilania, and R. Ramprasad, "Critical assessment of regression-based machine learning methods for polymer dielectrics," Comput. Mater. Sci. **125**, 123–135 (2016).
[53]O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," Nat. Commun. **8**, 15679 (2017).
[54]A. Mansouri Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, and J. Brgoch, "Machine learning directed search for ultraincompressible, superhard materials," J. Am. Chem. Soc. **140**, 9844–9853 (2018).
[55]I. Petousis, W. Chen, G. Hautier, T. Graf, T. D. Schladt, K. A. Persson, and F. B. Prinz, "Benchmarking density functional perturbation theory to enable high-throughput screening of materials for dielectric constant and refractive index," Phys. Rev. B **93**, 115151 (2016).
[56]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. **10**, 5316 (2019).
[57]L. Pauling, *The Nature of the Chemical Bond*, 3rd ed. (Cornell University Press, Ithaca, NY, 1960).
[58]J. C. Phillips and J. A. Van Vechten, "Spectroscopic analysis of cohesive energies and heats of formation of tetrahedrally coordinated semiconductors," Phys. Rev. B **2**, 2147–2160 (1970).